

Volume 9 Issue 8

August 2018



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor...

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 9 Issue 8 August 2018
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)
©2013 The Science and Information (SAI) Organization

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University

Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation

Associate Editors

Chao-Tung Yang

Department of Computer Science, Tunghai University, Taiwan

Domain of Research: Software Engineering and Quality, High Performance Computing, Parallel and Distributed Computing, Parallel Computing

Elena SCUTELNICU

"Dunarea de Jos" University of Galati, Romania

Domain of Research: e-Learning, e-Learning Tools, Simulation

Krassen Stefanov

Professor at Sofia University St. Kliment Ohridski, Bulgaria

Domains of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, Big Data, Cloud Computing, Data Retrieval and Data Mining, Distributed Systems, e-Learning Organisational Issues, e-Learning Tools, Educational Systems Design, Human Computer Interaction, Internet Security, Knowledge Engineering and Mining, Knowledge Representation, Ontology Engineering, Social Computing, Web-based Learning Communities, Wireless/ Mobile Applications

Maria-Angeles Grado-Caffaro

Scientific Consultant, Italy

Domain of Research: Electronics, Sensing and Sensor Networks

Mohd Helmy Abd Wahab

Universiti Tun Hussein Onn Malaysia

Domain of Research: Intelligent Systems, Data Mining, Databases

T. V. Prasad

Lingaya's University, India

Domain of Research: Intelligent Systems, Bioinformatics, Image Processing, Knowledge Representation, Natural Language Processing, Robotics

Reviewer Board Members

Aakash Ahmad Abbasi	Ali Asghar Pourhaji Kazem	Aris Skander Skander
Abbas Karimi	Ali Hamzeh	Arun D Kulkarni
Abbas M. Al-Ghaili	Ali Ismail Awad	Arun Kumar Singh
Abdelghni Lakehal	Ali Mustafa Qamar	Arvind K Sharma
Abdul Aziz Gill	Alicia Menchaca Valdez	Asadullah Shaikh
Abdul Hamid Mohamed Ragab	Altat Mukati	Asfa Praveen
Abdul Karim Assaf ABED	Aman Chadha	Ashok Matani
Abdul Razak	Amin Ahmad Shaqrah	Ashraf Hamdy Owis
Abdul Wahid Ansari	amine baina	ASIM TOKGOZ
Abdur Rashid Khan	Amir HAJJAM EL HASSANI	Asma Cherif
Abeer Mohamed ELkorany	Amirrudin Kamsin	Asoke Nath
ABRAHAM VARGHESE	Amitava Biswas	Athanasios Koutras
Adebayo Omotosho	Amjad Gawanmeh	Ayad Ghany Ismaeel
ADEMOLA ADESINA	Anand Nayyar	Ayman EL-SAYED
Aderemi A. Atayero	Anandhi Mohanraj Anu	Ayman Shehata Shehata
Adi A. Maaita	Andi Wahyu Rahardjo Emanuel	Ayoub BAHNASSE
Adnan Ahmad	Anews Samraj	Ayush Singhal
Adrian Nicolae Branga	Anirban Sarkar	Azam Moosavi
Ahmad A. Al-Tit	Anita Sofia V S	Babatunde Opeoluwa Akinkunmi
Ahmad A. Saifan	Anju Bhandari Gandhi	Bae Bossoufi
Ahmad Hoirul Basori	Anouar ABTOY	Balasubramanie Palanisamy
Ahmad Mousa Altamimi	Anshuman Sahu	BASANT KUMAR VERMA
Ahmed Boutejdar	Anthony Nosike Isizoh	Basem M. ElHalawany
Ahmed Nabih Zaki Rashed	Antonio Dourado	Basil Hamed
Ahmed S.A AL-Jumaily	Antonio Formisano	Basil M Hamed
Ahmed Z. Emam	ANUAR BIN MOHAMED KASSIM	Basim Almayahi
Ajantha Herath	Anuj Kumar Gupta	Bestoun S. Ahmed
Akram Belghith	Anuranjan misra	Bhanu Kaushik
Alaa F. Sheta	Appasami Govindasamy	Bhanu Prasad Pinnamaneni
Albert Alexander S	Arash Habibi Lashkari	Bharti Waman Gawali
Alci-nia Zita Sampaio	Aree Ali Mohammed	Bilian Song
Alexane Bouënard	Arfan Jaffar	Binod Kumar
ALI AMER ALWAN	ARINDAM SARKAR	Bogdan Belean

Bohumil Brtnik	Divya Kashyap	George D. Pecherle
Bouchaib CHERRADI	Djilali IDOUGHI	George Mastorakis
Brahim Raouyane	Dong-Han Ham	Georgios Galatas
Branko Karan	Dragana Becejski-Vujaklija	Gerard Dumancas
Bright Keswani	Duck Hee Lee	Ghalem Belalem Belalem
Brij Gupta	Duy-Huy NGUYEN	gherabi noredine
C Venkateswarlu Venkateswarlu Sonagiri	Ehsan Mohebi	Giacomo Veneri
Chanashekhhar Meshram	El Sayed A. Mahmoud	Giri Babu
Chao Wang	Elena Camossi	Goraksh Vithalrao Garje
Chao-Tung Yang	Elena SCUTELNICU	Govindarajulu Salendra
Charlie Obimbo	Elyes Maherzi	Grebenisan Gavril
Chee Hon Lew	Eric Tutu Tchao	Grigoras N. Gheorghe
CHERIF Med Adnen	Eui Chul Lee	Guandong Xu
Chien-Peng Ho	Evgeny Nikulchev	Gufran Ahmad Ansari
Chun-Kit (Ben) Ngan	Ezekiel Uzor OKIKE	Gunaseelan Devaraj
Ciprian Dobre	Fabio Mercorio	GYÖRÖDI ROBERT STEFAN
Constantin Filote	Fadi Safieddine	Hadj Hamma Tadjine
Constantin POPESCU	Fahim Akhter	Haewon Byeon
CORNELIA AURORA Gyorödi	Faizal Khan	Haibo Yu
Cosmina Ivan	FANGYONG HOU	Haiguang Chen
Cristina Turcu	Faris Al-Salem	Hamid Ali Abed AL-Asadi
Dai-Gyoung Kim	fazal wahab karam	Hamid Mukhtar
Daniel Filipe Albuquerque	Firkhan Ali Hamid Ali	Hamidullah Binol
Daniel Ioan Hunyadi	Fokrul Alom Mazarbhuiya	Hanan Elazhary
Daniela Elena Popescu	Fouad AYOUB	hanan habbi
Danijela Efnusheva	Francesco FP Perrotta	Hany Kamal Hassan
Dariusz Jakóbczak	Frank AYO Ibikunle	Harco Leslie Hendric SPITS WARNARS
Deepak Garg	Fu-Chien Kao	HARDEEP SINGH
Devena Prasad	G R Sinha	Hariharan Shanmugasundaram
DHAYA R	Gahangir Hossain	Harish Garg
Dheyaa Kadhim	Galya Nikolova Georgieva- Tsaneva	Hazem I. El Shekh Ahmed I. El Shekh Ahmed
Diaa Salama Dr	Gamil Abdel Azim	Heba Mahmoud Afify
Dimitris Chrysostomou	Ganesh Chandra Deka	Hela Mahersia
Dinesh Kumar Saini	Ganesh Chandra Sahoo	Hemalatha SenthilMahesh
Dipti Durgesh Patil	Gaurav Kumar	

Hesham G. Ibrahim	John P Sahlin	LATHA RAJAGOPAL
Hikmat Ullah Khan	JOHN S MANOHAR	Lazar Vojislav Stošić
Himanshu Aggarwal	JOSE LUIS PASTRANA	Le Li
Hongda Mao	José Santos Reyes	Leanos A Maglaras
Hossam Faris	Jui-Pin Yang	Leon Andretti Abdillah
Huda K. Kadhim AL-Jobori	Jungu J Choi	Lijian Sun
Hui Li	Jyoti Chaudhary	Liming Luke Chen
Hüseyin Oktay ERKOL	Jyoti Gautam	Ljubica B. Kazi
Ibrahim Adepoju Adeyanju	K V.L.N.Acharyulu	Ljubomir Jerinic
Ibrahim Missaoui	Ka-Chun Wong	Lokesh Kumar Sharma
Ikvinderpal Singh	Kamatchi R	Long Chen
Ilayaraja Muthalagu	Kamran Kowsari	M A Rabbani
Imad Zeroual	KANNADHASAN SURIYAN	M. Reza Mashinchii
Imed JABRI	KARTHIK MURUGESAN	M. Tariq Banday
Imran Ali Chaudhry	KASHIF MUNIR	Madiah Mohd Saudi
Imran Memon	Kashif Nisar	madjid khalilian
IRFAN AHMED	Kato Mivule	Mahdi H. Miraz
ISMAIL YUSUF	Kayhan Zrar Ghafoor	Mahmoud M Abd Ellatif
iss EL OUADGHIRI	Kennedy Chinedu Okafor	Mahtab Jahanbani Fard
Iwan Setyawan	KHAIRULLAH KHAN KHAN	Majharoddin Kazi Kazi
Jabar H Yousif	Khaled Loukhaoukha	majzoob kamal aldein omer
Jacek M. Czerniak	Khalid Mahmood	Malack Omae Oteri
Jafar Ahmad Alzubi	Khalid Nazim Sattar Abdul	Malik Muhammad Saad Missen
Jai Singh W	Khin Wee Lai	Mallikarjuna Reddy Doodipala
JAMAIAH HAJI YAHAYA	Khurram Khurshid	Man Fung LO
James Patrick Henry Coleman	KIRAN SREE POKKULURI	Manas deep
Jamil Abdulhamid Mohammed Saif	KITIMAPORN CHOOCHOTE	Manisha Gupta
Jatinderkumar Ramdass Saini	Kohei Arai	Manju Kaushik
Javed Anjum Sheikh	Kottakkaran Sooppy Nisar	Manmeet Mahinderjit Singh
Jayapandian N	kouki Mohamed	Manoharan P.S.
Jayaram M A	Krasimir Yankov Yordzhev	Manoj Manoj Wadhwa
Jerwinprabu A	Krassen Stefanov Stefanov	Manpreet Singh Manna
Ji Zhu	Krishna Kishore K V	Manuj Darbari
Jia Uddin Jia	Krishna Prasad Miyapuram	Marcellin Julius Antonio Nkenlifack
Jim Jing-Yan Wang	Labib Francis Gergis	Marek Reformat
	Lalit Garg	Maria-Angeles Grado-Caffaro

Marwan Alseid	Mohammed Shamim Kaiser	Naseer Ali Alquraishi
Mazin S. Al-Hakeem	Mohammed Tawfik Hussein	Nasrollah Pakniat
Md Ruhul Islam	Mohd Ashraf Ahmad	Natarajan Subramanyam
Md. Al-Amin Bhuiyan	Mohd Helmy Abd Wahab	Natheer Gharaibeh
Mehdi Bahrami	Mokhtar Beldjehem	Nayden V. Nenkov
Mehdi Neshat	Mona Elshinawy	Nazeeh Ghatasheh
Messaouda AZZOUZI	Monir Kaid	Nazeeruddin Mohammad
Milena Bogdanovic	Mostafa Mostafa Ezziyyani	Neeraj Kumar Tiwari
Miriampally Venkata Raghavendra	Mouhammad sharari sharari alkasassbeh	NEERAJ SHUKLA
Mirjana Popovic	Mounir Hemam	Nestor Velasco-Bermeo
Miroslav Baca	Mourad Amad	Nguyen Thanh Binh
Moamin Mahmoud	Mudasir Manzoor Kirmani	Nidhi Arora
Moeiz Miraoui	Mueen Uddin	NILAMADHAB MISHRA
Mohamed AbdelNasser	Muhammad Adnan Khan	Nilanjan Dey
Mohamed Mahmoud	Muhammad Abdul Rehman	Ning Cai
Mohamed Salah SALHI	Muhammad Asif Khan	Niraj Singhal
Mohamed A. El-Sayed	Muhammad Hafidz Fazli Bin Md Fauadi	Nithyanandam Subramanian
Mohamed Abdel Fatah Ashabrawy	Muhammad Naeem	Nizamud Din
Mohamed Ali Mahjoub	Muhammad Saeed	Noura Aknin
Mohamed Eldosoky	Muniba Memon	Obaida M. Al-Hazaimeh
Mohamed Hassan Saad Kaloup	MUNTASIR AL-ASFOOR	Olawande Justine Daramola
Mohamed Najeh LAKHOUA	Murphy Choy	Oliviu Matei
Mohamed SOLTANE Mohamed	Murthy Sree Rama Chandra Dasika	Om Prakash Sangwan
Mohammad Abdul Qayum	MUSLIHAH WOOK	Omaima Nazar Al-Allaf
Mohammad Ali Badamchizadeh	Mustapha OUJAOURA	Omar A. Alzubi
Mohammad Azzeh	MUTHUKUMAR S SUBRAMANYAM	Omar S. Gómez
Mohammad H. Alomari	N.Ch. Sriman Narayana Iyengar	Osama Ali Awad
Mohammad Haghighat	Nadeem Akhtar	Osama Omer
Mohammad Jannati	nafiul alam siddique	Ouchtati Salim
Mohammad Zarour	Nagy Ramadan Darwish	Ousmane THIARE
Mohammed Abdulhameed Al- shabi	Najeed Ahmed Khan	P.V. Praveen Sundar
Mohammed A. Akour	Najib A. Kofahi	Paresh V Virparia
Mohammed Ali Hussain	Namrata Dhanda	Parminder Singh Kang
Mohammed Sadgal	Nan Wang	PAUL CELICOURT
		Peng Xia
		Ping Zhang

Piyush Kumar Pareek	Reza Fazel-Rezai	Senol Piskin
Poonam Garg	Reza Ghasemy Yaghin Dr Reza Ghasemy Yaghin	SENTHIL P Prof
Prabhat K Mahanti	Riaz Ul-Amin	Sérgio André Ferreira
PRASUN CHAKRABARTI	Ricardo Ângelo Rosa Vardasca	Seyed Hamidreza Mohades Kasaei
Praveen Kumar	Ritaban Dutta	Shadi Mahmoud Atalla
PRISCILLA RAJADURAI	Rodica Doina Zmaranda	Shafiqul Abidin
PROF DURGA PRASAD SHARMA (PHD)	Rohini Ravi	Shahab Shamshirband
Purwanto Purwanto	Rohit Raja	Shahanawaj Ahamad
Qaisar Abbas	Roopali Garg	Shaidah Jusoh
Qifeng Qiao	roslina ibrahim	Shaiful Bakri Ismail
Rachid Saadane	Ruchika Malhotra	Shailesh Kumar
Radwan R. Tahboub	Rutvij H. Jhaveri	Shakir Gayour Khan
raed Kanaan	SAADI Slami	Shashi Dahiya
Raghuraj Singh	Sachin Kumar Agrawal	Shawki A. Al-Dubae
Rahul Malik	Sagarmay Deb	Sheeraz Ahmed Dr.
Raja Ramachandran	Sahar Abd EL_RAhman Ismail	Sheikh Ziauddin
raja sarath kumar boddu	Said Ghoniemy	Sherif E. Hussein
Rajesh Kumar	Said Jadid Abdulkadir	Shishir Kumar
Rakesh Chandra Balabantaray	Sajal Bhatia	SHOBA MOHAN
Rakesh Kumar Dr.	Saman Hina	Shriniwas Vasantrao Chavan
Ramadan Elaiess	SAMSON OLUWASEUN FADIYA	Shriram K Vasudevan
Ramani Kannan	Sanam Shahla Rizvi	Siddeeq Ameen
RAMESH MUTHUSAMY	Sandeep R Reddivari	Siddhartha Jonnalagadda
RAMESH VAMANAN	Sangeetha SKB	Sim-Hui Tee
Rana Khudhair Abbas Ahmed	Sanskruti V Patel	Simon L. R. Vrhovec
Rashad Abdullah Al-Jawfi	Santosh Kumar	Simon Uzezi Ewedafe
Rashid Sheikh	Sasan Adibi	Siniša Opic
Ratnesh Litoriya	Sattar Bader Sadkhan	Sivakumar Poruran
Ravi Kiran Varma P	Satyena Prasad Singh	sivaranjani reddy
Ravi Prakash	Sebastian Marius Rosu	Slim BEN SAOUD
RAVINA CHANGALA	Secui Dinu Calin	Sobhan Roshani
Ravisankar Hari	Seema Shah	Sofien Mhatli
Rawya Y. Rizk	Seifedine Nimer Kadry	sofyan Mohammad Hayajneh
Rayed AlGhamdi	Selem Charfi	Sohail Jabbar
Reshmy Krishnan	SENGOTTUVELAN P	Sri Devi Ravana

Sudarson Jena	Taskeed Jabid	Wenzhao Zhang
Sudipta Roy	Tasneem Bano Rehman	Wichian Sittiprapaporn
Suhail Sami Owais Sami Owais Owais	thabet Mohamed slimani	Xi Zhang
Suhas J Manangi	Totok R. Biyanto	Xiao Zhang
SUKUMAR SENTHILKUMAR	Touati Youcef	Xiaojing Xiang
Süleyman Eken	Tran Xuan Sang	Xiaolong Wang
Sumazly Sulaiman	TSUNG-CHUAN MA	Xunchao Hu
Sumit Goyal	Tsvetanka Georgieva-Trifonova	Y Srinivas
Sunil Phulre	Uchekukwu Awada	Yanping Huang
Suparerk Janjarasjitt	Udai Pratap Rao	Yao-Chin Wang
Suresh Sankaranarayanan	Urmila N Shrawankar	Yasser M. Alginahi
Surya Narayan Panda	V Baby Deepa	Yaxin Bi
Susarla Venkata Ananta Rama Sastry	Vaidas Giedrimas	Yi Fei Wang
Suseendran G	Vaka MOHAN	YI GU
Suxing Liu	Venkata Raghavendran Chaluvadi	Yihong Yuan
Syed Asif Ali	VENKATESH JAGANATHAN	Yilun Shang
T C.Manjunath	Vijay Bhaskar Semwal	Yu Qi
T V Narayana rao Rao	Vijayarani Mohan S	Zacchaeus Oni Omogbadegun
T. V. Prasad	Vijendra Singh	Zaffar Ahmed Shaikh
Taghi Javdani Gandomani	Vinayak K Bairagi	Zairi Ismael Rizman
Taiwo Ayodele	VINCE PAUL A	Zarul Fitri Zaaba
Talal Bonny	Visara Urovi	Zeki Yetgin
Tamara Zhukabayeva	Vishnu Narayan Mishra	Zenzo Polite Ncube
Taner Tuncer	Vitus S.W. Lam	ZHENGYU YANG
Tanvi Banerjee	VNR SAIKRISHNA K	Zhigang Yin
Tanweer Alam	Voon Ching Khoo	Zhihan Lv
Tanzila Saba	VUDA SREENIVASARAO	Zhixin Chen
TAOUFIK SALEM SAIDANI	Wali Khan Mashwani	Zia Ur Rahman Zia
Tarek Fouad Gharib	Wei Wei	Ziyue Xu
tarig ahmed	Wei Zhong	Zlatko Stapic
	Wenbin Chen	Zne-Jung Lee
		Zuraini Ismail

CONTENTS

Paper 1: Framework Utilizing Machine Learning to Facilitate Gait Analysis as an Indicator of Vascular Dementia
Authors: Arshia Khan, Janna Madden, Kristine Snyder

PAGE 1 – 6

Paper 2: Recognition of Ironic Sentences in Twitter using Attention-Based LSTM
Authors: Andrianarisoa Tojo Martini, Makhmudov Farrukh, Hongwei Ge

PAGE 7 – 11

Paper 3: The Role of Camera Convergence in Stereoscopic Video See-through Augmented Reality Displays
Authors: Fabrizio Cutolo, Vincenzo Ferrari

PAGE 12 – 17

Paper 4: Comparison of Event Choreography and Orchestration Techniques in Microservice Architecture
Authors: Chaitanya K. Rudrabhatla

PAGE 18 – 22

Paper 5: Location-based E-Commerce Services: (Re-) Designing using the ISO9126 Standard
Authors: Antonia Stefani, Bill Vassiliadis, Theofanis Efthimiades

PAGE 23 – 33

Paper 6: Programming Technologies for the Development of Web-Based Platform for Digital Psychological Tools
Authors: Evgeny Nikulchev, Dmitry Ilin, Pavel Kolyasnikov, Vladimir Belov, Ilya Zakharov, Sergey Malykh

PAGE 34 – 45

Paper 7: Review of Prediction of Disease Trends using Big Data Analytics
Authors: Diellza Nagavci, Mentor Hamiti, Besnik Selimi

PAGE 46 – 50

Paper 8: The Role of Hyperspectral Imaging: A Literature Review
Authors: Muhammad Mateen, Junhao Wen, Nasrullah, Muhammad Azeem Akbar

PAGE 51 – 62

Paper 9: A Review on Scream Classification for Situation Understanding
Authors: Saba Nazir, Muhammad Awais, Sheraz Malik, Fatima Nazir

PAGE 63 – 75

Paper 10: Performance Improvement of Web Proxy Cache Replacement using Intelligent Greedy-Dual Approaches
Authors: Waleed Ali

PAGE 76 – 85

Paper 11: El Niño / La Niña Identification based on Takens Reconstruction Theory
Authors: Kohei Arai, Kaname Seto

PAGE 86 – 91

Paper 12: Adaptive Return of e-Training (ROT) based on Communication Technology
Authors: Fahad Alotaibi

PAGE 92 – 97

Paper 13: Comparison of Hash Function Algorithms Against Attacks: A Review

Authors: Ali Maetouq, Salwani Mohd Daud, Noor Azurati Ahmad, Nurazeen Maarop, Nilam Nur Amir Sjarif, Hafiza Abas

PAGE 98 – 103

Paper 14: An Analysis of Cloud Computing Adoption Framework for Iraqi e-Government

Authors: Ban Salman Shukur, Mohd Khanapi Abd Ghani, M.A. Burhanuddin

PAGE 104 – 112

Paper 15: A Survey on Tor Encrypted Traffic Monitoring

Authors: Mohamad Amar Irsyad Mohd Aminuddin, Zarul Fitri Zaaba, Manmeet Kaur Mahinderjit Singh, Darshan Singh Mahinder Singh

PAGE 113 – 120

Paper 16: The Implementation of Computer based Test on BYOD and Cloud Computing Environment

Authors: Ridi Ferdiana, Obert Hoseanto

PAGE 121 – 124

Paper 17: Method for Designing Scalable Microservice-based Application Systematically: A Case Study

Authors: Ahmad Tarmizi Abdul Ghani, Mohamad Shanudin Zakaria

PAGE 125 – 135

Paper 18: Adaptive Simulated Evolution based Approach for Cluster Optimization in Wireless Sensor Networks

Authors: Abdulaziz Alsayyari

PAGE 136 – 143

Paper 19: Investigating the Acceptance of Mobile Health Application User Interface Cultural-Based Design to Assist Arab Elderly Users

Authors: Ahmed Alsswey, Irfan Naufal Bin Umar, Brandford Bervell

PAGE 144 – 152

Paper 20: Acoustic Classification using Deep Learning

Authors: Muhammad Ahsan Aslam, Muhammad Umer Sarwar, Muhammad Kashif Hanif, Ramzan Talib, Usama Khalid

PAGE 153 – 159

Paper 21: Quality Assurance for Data Analytics

Authors: Rakesh Kumar, Birth Subhash, Maria Fatima, Waqas Mahmood

PAGE 160 – 166

Paper 22: Developing Communication Strategy for Multi-Agent Systems with Incremental Fuzzy Model

Authors: Sam Hamzeloo, Mansoor Zolghadri Jahromi

PAGE 167 – 174

Paper 23: OpenSimulator based Multi-User Virtual World: A Framework for the Creation of Distant and Virtual Practical Activities

Authors: MOURDI Youssef, SADGAL Mohamed, BERRADA FATHI Wafaa, EL KABTANE Hamada

PAGE 175 – 186

Paper 24: Performance Evaluation of Cloud Computing Resources

Authors: Muhammad Sajjad, Arshad Ali, Ahmad Salman Khan

PAGE 187 – 199

Paper 25: Intelligent Model Conception Proposal for Adaptive Hypermedia Systems

Authors: Mehdi TMIMI, Mohamed BENSLIMANE, Mohammed BERRADA, Kamar OUZZANI

PAGE 200 – 205

Paper 26: Cost Aware Resource Selection in IaaS Clouds

Authors: Uzma Bibi

PAGE 206 – 214

Paper 27: ECG Abnormality Detection Algorithm

Authors: Soha Ahmed, Ali Hilal-Alnaqbi, Mohamed Al Hemaury, Mahmoud Al Ahmad

PAGE 215 – 219

Paper 28: Efficient Resource Consumption by Dynamic Clustering and Optimized Routes in Wireless Sensor Networks

Authors: Farzad Kiani

PAGE 220 – 226

Paper 29: A Method of Automatic Domain Extraction of Text to Facilitate Retrieval of Arabic Documents

Authors: Mohammad Khaled A. Al-Maghasbeh, Mohd Pouzi bin Hamzah

PAGE 227 – 230

Paper 30: Features and Potential Security Challenges for IoT Enabled Devices in Smart City Environment

Authors: Gasim Alandjani

PAGE 231 – 238

Paper 31: Comparative Study of PMSG Controllers for Variable Wind Turbine Power Optimization

Authors: Asma Hammami, Imen Saidi, Dhaou Soudani

PAGE 239 – 246

Paper 32: Impact and Challenges of Requirements Management in Enterprise Resource Planning (ERP) via ERP Thesaurus

Authors: Rahat Izhar, Dr. Shahid Nazir Bhatti, Saba Izhar, Dr. Amr Mohsen Jadi

PAGE 247 – 258

Paper 33: Implementation of Blended Learning in Teaching at the Higher Education Institutions of Pakistan

Authors: Saira Soomro, Arjumand Bano Soomro, Tariq Bhatti, Najma Imtiaz Ali

PAGE 259 – 264

Paper 34: The Measurement of Rare Plants Learning Media using Backward Chaining Integrated with Context-Input-Process-Product Evaluation Model based on Mobile Technology

Authors: Nyoman Wijana, Ni Nyoman Parmithi, I Gede Astra Wesnawa, I Made Ardana, I Wayan Eka Mahendra, Dewa Gede Hendra Divayana

PAGE 265 – 277

Paper 35: Energy Consumption Evaluation of AODV and AOMDV Routing Protocols in Mobile Ad-Hoc Networks

Authors: Fawaz Mahiuob Mohammed Mokbal, Khalid Saeed, Wang Dan

PAGE 278 – 288

Paper 36: Piezoelectric based Biosignal Transmission using Xbee

Authors: Mohammed Jalil, Mohamed Al Hamadi, Abdulla Saleh, Omar Al Zaabi, Soha Ahmed, Walid Shakhathreh, Mahmoud Al Ahmad

PAGE 289 – 294

Paper 37: Performance Evaluation of a Smart Remote Patient Monitoring System based Heterogeneous WSN

Authors: Mohamed EDDABBAH, Mohamed MOUSSAOUI, Yassin LAAZIZ

PAGE 295 – 300

Paper 38: Mapping Wheat Crop Phenology and the Yield using Machine Learning (ML)

Authors: Muhammad Adnan, Abaid-ur-Rehman, M. Ahsan Latif, Naseer Ahmad, Maria Nazir, Naheed Akhter

PAGE 301 – 306

Paper 39: Soft Error Tolerance in Memory Applications

Authors: Muhammad Sheikh Sadi, Md. Shamimur Rahman, Shaheena Sultana, Golam Mezbah Uddin, Kazi Md. Bodrul Kabir

PAGE 307 – 314

Paper 40: Safety and Performance Evaluation Method for Wearable Artificial Kidney Systems

Authors: YeJi Ho, SangHoon Park, KyungMin Jo, Barum Choi, SangEun Park, Jaesoon Choi

PAGE 315 – 319

Paper 41: Data Mining Models Comparison for Diabetes Prediction

Authors: Amina Azrar, Yasir Ali, Muhammad Awais, Khurram Zaheer

PAGE 320 – 323

Paper 42: Using Artificial Intelligence Approaches to Categorise Lecture Notes

Authors: Naushine Bibi Baijoo, Khusboo Bharossa, Somveer Kishnah, Sameerchand Pudaruth

PAGE 324 – 328

Paper 43: EEG-Based Emotion Recognition using 3D Convolutional Neural Networks

Authors: Elham S. Salama, Reda A.El-Khoribi, Mahmoud E.Shoman, Mohamed A.Wahby Shalaby

PAGE 329 – 337

Paper 44: Enhanced and Improved Hybrid Model to Prediction of User Awareness in Agriculture Sector

Authors: A.V.S. Pavan Kumar, Dr. R. Bhramaramba

PAGE 338 – 343

Paper 45: Identifying Dynamic Topics of Interest across Social Networks

Authors: Mohamed Salaheldin Aly, Abeer Al Korany

PAGE 344 – 349

Paper 46: Processing Sampled Big Data

Authors: Arshia Khan, Janna Madden, Kristine Snyder

PAGE 350 – 356

Paper 47: Access Control Model for Modern Virtual e-Government Services: Saudi Arabian Case Study

Authors: Rand Albrahim, Hessah Alsalamah, Shada Alsalamah, Mehmet Aksoy

PAGE 357 – 364

Paper 48: Evaluation of the Impact of Usability in Arabic University Websites: Comparison between Saudi Arabia and the UK

Authors: Mohamed Benaida, Abdallah Namoun, Ahmad Taleb

PAGE 365 – 375

Paper 49: Using Sab-lomha for an Alpha Channel based Image Forgery Detection

Authors: Muhammad Shahid Bhatti, Syed Asad Hussain, Abdul Qayyum, Abdul Karim Shahid, Muhammad Usman Akram and Sajid Ibrahim Hashmi

PAGE 376 – 384

Paper 50: Recommendations for Building Adaptive Cognition-based E-Learning

Authors: Mostafa Saleh, Reda Mohamed Salama

PAGE 385 – 393

Paper 51: Segmentation Method for Pathological Brain Tumor and Accurate Detection using MRI

Authors: Khurram Ejaz, Mohd Shafry Mohd Rahim, Amjad Rehman, Huma Chaudhry, Tanzila Saba, Anmol Ejaz, Chaudhry Farhan Ej

PAGE 394 – 401

Paper 52: Skew Detection and Correction of Mushaf Al-Quran Script using Hough Transform

Authors: Salem Saleh Bajfaish, Mohd Sanusi Azmi, Mohammed Nasser Al-Mhiqani, Amirul Ramzani Radzid, Hairulnizam Mahdin

PAGE 402 – 409

Paper 53: Review of Information Security Policy based on Content Coverage and Online Presentation in Higher Education

Authors: Arash Ghazvini, Zarina Shukur, Zaihosnita Hood

PAGE 410 – 423

Paper 54: Implementation of a Formal Software Requirements Ambiguity Prevention Tool

Authors: Rasha Alomari, Hanan Elazhary

PAGE 424 – 432

Paper 55: A Comparative Study of the Decisional Needs Engineering Approaches

Authors: OUTFAROUIN Ahmad, ZAHID Nouredine, ABDALI Abdelmounaïm

PAGE 433 – 441

Paper 56: A Blockchain Technology Evolution Between Business Process Management (BPM) and Internet-of-Things (IoT)

Authors: Doaa Mohey El-Din M. Hussein, Mohamed Hamed N. Taha, Nour Eldeen M. Khalifa

PAGE 442 – 450

Paper 57: Defects Prediction and Prevention Approaches for Quality Software Development

Authors: Mashooque Ahmed Memon, Mujeeb-Ur-Rhman Magsi Baloch, Muniba Memon, Syed Hyder Abbas Musavi

PAGE 451 – 457

Paper 58: Design and Implementation of a Risk Management Tool: A Case Study of the Moodle Platform

Authors: Nadia Chafiq, Mohammed Talbi, Mohamed Ghazouani

PAGE 458 – 461

Paper 59: Artificial Neural Network based Weather Prediction using Back Propagation Technique

Authors: Saboor Ahmad Kakar, Naveed Sheikh, Adnan Naseem, Saleem Iqbal, Abdul Rehman, Aziz ullah Kakar, Bilal Ahmad Kakar, Hazrat Ali Kakar, Bilal Khan

PAGE 462 – 470

Paper 60: An Incremental Technique of Improving Translation

Authors: Aasim Ali, Arshad Hussain

PAGE 471 – 474

Paper 61: Role Term-Based Semantic Similarity Technique for Idea Plagiarism Detection

Authors: Ahmed Hamza Osman, Hani Moetque Aljahdali

PAGE 475 – 484

Paper 62: Impact of Security in QoS Signaling in NGN: Registration Study

Authors: RAOUYANE Brahim, BELMEKKI Elmostafa, KHAIRI sara, BELLAFKIH mostafa

PAGE 485 – 492

Paper 63: Information System Quality: Managers Perspective

Authors: Sarah Aouhassi, Mostafa Hanoune

PAGE 493 – 502

Paper 64: Using Fuzzy Clustering Powered by Weighted Feature Matrix to Establish Hidden Semantics in Web Documents

Authors: Pramod D Patil, Parag Kulkarni

PAGE 503 – 514

Paper 65: A New E-Health Tool for Early Identification of Voice and Neurological Pathologies by Speech Processing

Authors: Bouafif Lamia, Ellouze Noureddine

PAGE 515 – 522

Paper 66: An Overview of Mutation Strategies in Bat Algorithm

Authors: Waqas Haider Bangyal, Jamil Ahmad, Hafiz Tayyab Rauf, Sobia Pervaiz

PAGE 523 – 534

Paper 67: Arabic Chatbots: A Survey

Authors: Sarah AlHumoud, Asma Al Wazrah, Wafa Aldamegh

PAGE 535 – 541

Paper 68: Learner Cognitive Behavior and Influencing Factors in Web-based Learning Environment

Authors: Kalla Madhusudhana

PAGE 542 – 546

Paper 69: New Hybrid Task Scheduling Algorithm with Fuzzy Logic Controller in Grid Computing

Authors: Younes Hajoui, Omar Bouattane, Mohamed Youssfi, Elhocein Illoussamen

PAGE 547 – 554

Paper 70: Performance Comparison of QEC Network based JAVA Application and Web based PHP Application

Authors: Sanaullah Memon, Rasool Bux Palh, Muniba Memon, Hina Siddique Memon

PAGE 555 – 564

Paper 71: Framework Utilizing Machine Learning to Facilitate Gait Analysis as an Indicator of Vascular Dementia

Authors: Cristian Vidal Silva, Rodolfo Villarroel, José ´Rubio, Franklin Johnson, ´Erika Madariaga, Alberto Urz´ua, Luis Carter, Camilo Campos-Vald´es, Xaviera A. L´opez-Cort´es

PAGE 565 – 574

Paper 72: A New Message Encryption Method based on Amino Acid Sequences and Genetic Codes

Authors: Ahmed Mahdee Abdo, Adel Sabry Essa, Abdullah A. Abdullah

PAGE 575 – 578

Paper 73: Using Artificial Neural Networks for Detecting Damage on Tobacco Leaves Caused by Blue Mold

Authors: Himer Avila-George, Topacio Valdez-Morones, Humberto P´erez-Espinosa, Brenda Acevedo-Ju´arez, Wilson Castro

PAGE 579 – 583

Paper 74: Complex Shear Modulus Estimation using Integration of LMS/AHI Algorithm

Authors: Quang-Hai Luong, Manh-Cuong Nguyen, TonThat-Long, Duc-Tan Tran

PAGE 584 – 589

Paper 75: RASP-TMR: An Automatic and Fast Synthesizable Verilog Code Generator Tool for the Implementation and Evaluation of TMR Approach

Authors: Abdul Rafay Khatri, Ali Hayek, and Josef Borcsok

PAGE 590 – 597

Paper 76: Design of Linear Time Varying Flatness-Based Control for Single-Input Single-Output Systems

Authors: Marouen Sleimi, Mohamed Ben Abdallah, Mounir Ayadi

PAGE 598 – 604

Paper 77: Comparative Performance Analysis of Efficient MIMO Detection Approaches

Authors: Muhammad Faisal, Fazal Wahab Karam, Ali Zahir, Sajid Bashir

PAGE 605 – 615

Paper 78: Sentiment Analysis, Visualization and Classification of Summarized News Articles: A Novel Approach

Authors: Siddhaling Urologin

PAGE 616 – 625

Paper 79: Minimization of Information Asymmetry Interference using Partially Overlapping Channel Allocation in Wireless Mesh Networks

Authors: Sadiq Shah, Khalid Saeed, Mustafa Khan, Rafi Ullah Khan, Mohib Ullah Khan, Misbah Daud Arbab Waseem Abbas

PAGE 626 – 634

Paper 80: Initialization Method for Communication and Data Sharing in P2P Environment Between Wireless Sensor Nodes

Authors: M. Asif Jamal, Aziz Ur Rehman, Moonisa Ahsan, M. S. Riaz, M. S. Zafar

PAGE 635 – 640

Paper 81: Formal Specification of Memory Coherence Protocol

Authors: Jahanzaib Khan, Muhammad Afiif, Muhammad Khurram Zahoor Bajwa, Muhammad Sohaib Mahmood, Sobia Usman

PAGE 641 – 650

Paper 82: Digital Technology Disorder: Justification and a Proposed Model of Treatment

Authors: Andrew Kear, Sasha L. Folkes

PAGE 651 – 665

Framework Utilizing Machine Learning to Facilitate Gait Analysis as an Indicator of Vascular Dementia

Arshia Khan, Janna Madden

Department of Computer Science, University of Minnesota,
Duluth
Duluth, Minnesota, United States

Kristine Snyder

Department of Mathematics and Statistics
University of Minnesota, Duluth
Duluth, Minnesota, United States and Stryd,
Boulder, Colorado, United States

Abstract—Vascular dementia (VD), the second most common type of dementia, affects approximately 13.9 per cent of people over the age of 71 in the United States alone. 26% of individuals develop VD after being diagnosed with congestive heart failure. Memory and cognition are increasingly affected as dementia progresses. However, these are not the first symptoms to appear in some types of dementia. Alterations in gait and executive functioning have been associated with Vascular Cognitive Impairment (VCI). Research findings suggest that gait may be one of the earliest affected systems during onset of VCI, immediately following a vascular episode. The diagnosis tools currently utilized for VD are focused on memory impairment, which is only observed in later stages of VD. Hence we are proposing a framework that isolates gait and executive functioning analysis by applying machine learning to predict VD before cognition is affected, so pharmacological treatments can be used to postpone the onset of cognitive impairment. Over a period of time, we hope to be able to develop prediction algorithms that will not only identify but also predict vascular dementia.

Keywords—Gait; machine learning; vascular dementia; early diagnosis; indicators; gait analysis

I. INTRODUCTION

The effects of vascular dementia are widespread. Currently, approximately 13.9% of individuals over the age of 71 have some form of dementia. As the demographics of our population change, the impact of chronic disease continues to increase. Currently, there are 35 million people in the United States over the age of 65. By 2030, this number is expected to be more than 70 million [1]. Vascular Cognitive impairment (VCI), the second widely spread type of dementia [2], second only to Alzheimer's Dementia (AD), occurs in 26% of individuals who are diagnosed with Congestive Heart Failure (CHF) [3]. Dementia is of particular concern as the decline of memory and cognition functioning lead to a loss of independence and increased dependency on families and healthcare systems [1]. The prevalence of Vascular Dementia (VD), second to (AD), doubles every 5.3 years, while the prevalence of AD doubles every 4.3 year. Just like AD, the severity of VD is related to age. The age-adjusted rates of VCI are 14.6 per thousand per years, while it is 19.2 for AD [4]. Much of the healthcare costs incurred from dementia fall into the realm of government programs. However, individual caregivers also experience decrease in income and chronic fatigue for their efforts [1]. Early diagnosis of dementia, cannot stop the effect of the disease. While, it has been

suggested that early treatment may slow the progression of the disease [5], developing an evaluation or monitoring protocol has proven challenging.

In this paper, we look at the background of subcortical vascular cognitive impairment, the use of technology in its diagnosis, typical onset patterns and early indicators. Following this a framework for applying machine learning techniques to these early indicators is presented. The paper concludes with a discussion on current techniques and future research direction.

II. BACKGROUND

Vascular dementia, the second most common type of dementia [6]-[9] was first identified by Thomas Willis in 1684 as a palsy that would cause 'stupidity' after stroke. Approximately 20-30% of individuals affected by stroke develop vascular dementia [2]. 36 million people around the world currently suffer from dementia. Vascular occlusions of larger vessels result in stroke, whereas the occlusion of smaller vessels results in a gradual degradation of subcortical areas of the brain that perform executive functions. This condition of small vessel occlusion is termed as Subcortical Vascular Dementia, (SVD), a most common type of VD [6], [8], [9]. 36-67% of the vascular dementia is subcortical vascular dementia [10]. Our paper will focus on the SVD form of vascular dementia.

The use of technology as a tool in the medical setting is a growing area of research and development. With the desire to reduce healthcare costs and support patient-centered care, the integration of technology into medical processes is increasingly common [11]. At first these technologies were designed primarily for practitioners, managers, and other professionals in a healthcare setting. However, as telecommunication and mobile computing technologies improved, systems that interact directly with patients outside of the healthcare setting became more prevalent [12]. Perhaps the strongest force driving this innovation is the growing interest in evidence-based and personalized medicine. The origins of evidence-based medicine can be traced back to mid-19th century Paris where physicians "conscientiously and explicitly" used outcomes from their previous cases to make current care decisions [13]. The rise of technology has greatly changed how evidence-based medicine is practiced, but the core idea has remained the same -- incorporating clinical

evidence from systematic research into health decisions, particularly diagnosis [13]. Likewise, personalized medicine looks to base health decisions on individual health records, the idea being that individuals' patterns and health history should be incorporated into health decisions. Often referred to as N-of-1 studies, mobile technology has proven to be of great utility in the execution of these studies to collect, record and communicate current data with providers [14].

Personalized medicine tools have been applied to many realms of health, including preventive care, chronic disease management, and monitoring of patients. Such systems allow us to ask previously unanswerable questions to better understand the connection between behavioral choices and health. There are many benefits to such a system. However, they come with many challenges. Because of the sheer mass of data, it is necessary to develop analytical methods to process raw data into actionable knowledge for patients and their providers [15]. In addition, regulatory, financial reimbursement, and technical security hurdles need to be considered.

This research looks at the potential of machine learning tools to analyze gait. Research has begun to recognize patterns in vascular dementia onset. However, in many cases, particular elements of gait are studied in isolation. We are proposing a framework that incorporates a larger number of measurable elements of gait and utilizes machine learning to optimize and improve upon past work.

A. Progression of Vascular Dementia

Progression of Vascular Dementia generally represents a stepwise decline – appearing suddenly after episode and aggravated from following episodes, but without the continuous decline common to Alzheimer's Disease. Vascular Dementia transitions from preclinical to Vascular Cognitive Impairment to Vascular Dementia can be sub-classified as mild, moderate or severe. To understand each diagnosis, the symptoms and expected progression at each of these stages must be considered.

1) *Risk factors*: The risk factors for vascular dementia are the same as cardiovascular diseases such as hypertension, stroke, atrial fibrillation, aortic fibrillation, diabetes mellitus type 2, obesity, lack of active lifestyle, depression, sleep apnea, and smoking [6]. In addition, the hemiparetic type of gait was identified as a risk factor [16].

2) *Symptoms*: Two of the foremost cognitive domains affected by vascular dementia are gait and executive functioning [17]. Among other cognitive domains affected, such as problem solving, execution of complex commands and motor skills, memory is the most significant [6].

3) *Preclinical Stage*: The initial stage is often described as "silent", as the brain begins to change without measureable symptoms being displayed. Changes are not detectable on tests and the symptoms patient experiences are not diagnosable. Because of this, much that is known about the preclinical stage of Vascular Dementia is based on retrospective evaluations of records of diagnosed cases. One such study found that patients had memory complaints 12

years prior to diagnosis and had experienced declines in activities of daily living 5 to 7 years previous to diagnosis [18]. While Vascular Dementia patients had memory complaints 12 years prior to diagnosis, cognitively, there is comparatively less deterioration in the preclinical stage as compared to other forms of Dementia. Patients' with incident vascular dementia deteriorate earlier and faster in daily functioning, especially the more physical activities of daily living such as activities, arising, dressing and grooming, eating, hygiene, grip, reach, and walking, as compared to other forms of Dementia that experience the first changes in cognitive activities such as finance management, phoning, medication use, housekeeping, and meal preparation [18], [19]. In addition, the preclinical stage is often accompanied by symptoms of depression, particularly motivation-related such as lack of interest, loss of energy and concentration difficulties. This association still remained significant after adjusting for memory complaints, showing that depressive symptoms are not merely a by-product of perceived cognitive difficulties [20].

4) *Vascular Cognitive Impairment*: The progression from preclinical to Vascular Cognitive Impairment is not an unambiguous transition. The Vascular Cognitive Impairment stage is loosely defined as cases where one or more cognitive domains becomes significantly affected [21], [22]. At this stage in the disease, symptoms are becoming clinically detectable and while noticeable in daily living, they are not generally too limiting in this respect.

5) *Vascular Dementia*: Onset of Vascular Dementia is marked by cognitive impairment severe enough to interfere with everyday activities. The onset of Vascular Dementia can be divided into sub-domains of mild, moderate, moderately severe, and severe.

6) *Mixed Dementia*: Another factor to consider is the onset of other forms of cognitive impairment in addition to vascular dementia, referred to as "Mixed Dementia". Mixed Dementia refers to cases of Vascular Dementia where symptoms of other cognitive impairment, not originating from the vascular episode, begin to affect the patient in addition to the symptoms of Vascular Dementia already present. Approximately 15 percent of cases of Vascular Dementia present with other forms of cognitive impairment [23]. Identifying the onset of other cognitive impairments and the relationship between the Vascular episode and these other cognitive impairments is one of the great challenges facing research in this area.

B. Gait as an early Indicator of Vascular Dementia

The challenge of diagnosing Vascular Dementia during or prior to vascular cognitive impairment is one that much research has looked into. Vascular dementia in particular is of interest because patients often experience a "step-wise" deterioration as opposed to gradual and symptomatic decline following acute events. In addition, in the subcategory of vascular dementia, mild cognitive impairment seems equally if not more prevalent than full onset of dementia symptoms [24].

Researchers have identified several indicators of vascular cognitive impairment including impaired social or occupational functioning, motor activities, visual processing and abstract reasoning [24]. However, one challenge is to transform these rather abstract indicators into quantifiable and actionable metrics.

One developing area is that of gait analysis. This could be stride length, lateral balance, or effort exerted (measured using heart-rate monitor for example) for a particular class of activity [24]. Motor activity metrics focus on measuring ease, frequency, and type of movement. Gait has clear links to motor activities, but it also has an interesting link to visual processing since the visual system is strongly correlated with balance. In comparison to other physiological feedback systems, Visual information is typically more sensitive and is believed to play a significant role in fine-grained adjustments to balance, especially in the feet and ankles [25]. In patients with Vascular Dementia onset, impaired visual processing could be recognized in balance and gait discrepancies. Various gait metrics have been investigated, and their potential to identify vascular cognitive impairment has been evaluated.

1) *Distance and Speed Metrics*: In respect to gait analysis, distance and speed are typically measured in terms of a single stride. To understand what these metrics quantify, it's important to understand the components of a stride. Stride length is measured as the distance between two consecutive footfalls of the same foot. In addition, a single stride can be broken down into components, commonly, swing time and stance time. Swing time refers the time when only one foot is on the ground. The stance time refers to the time when that foot is on the ground. Distance and speed metrics are often in reference to either a complete stride or a component of a stride [26]. Shorter stride length has been linked to a increased chance of mild cognitive impairment [25]. Similarly, the a slower gait can indicate motor function concerns, thus measuring the velocity or stride frequency (steps per minute) can also provide useful information [26].

2) *Gait Disturbances and Difficulties*: In a similar vein, measuring gait disturbances or difficulties takes stride metrics and looks for abnormalities or inconsistencies between multiple instances over a given time period. In these methods, a range of acceptable metrics is defined and measurements outside that range are considered a disturbance [27]. This idea comes up frequently in fall detection, but the idea of creating a customized range based on user data rather than setting a predefined range, is one application of this idea that may prove useful to this research.

3) *Pace, Rhythm, and Variability of Gait*: Pace, rhythm, and variability of gait also look at comparing multiple strides in a time frame against each other. However, rather than looking for outliers, these metrics attempt to recognize patterns in the data and cases where these patterns are not true [27]. One example of this is measuring the percentage of the stride that is spent on the swing verses the percent on the stance. Typically, the swing takes 40 percent of the stride with stance taking the remaining 60 percent of the time [28]. Pace,

rhythm, and variability measures are aimed to quantify how much a gait changes in this respect over a time period.

4) *Stability*: Stability and width of base, while associated with pace, are largely metrics of balance. Stability looks at the consistency of weight patterns on a foot. Inconsistent (variable) patterns of weight distribution could indicate the individual is "shaky" and struggling to balance. Width of base is another way we can measure stability. Width of base refers to how far apart a individual's stance is. A wide stance can indicate balance concerns as well [28]. Finally, double support, when both feet are in contact with the ground at the same time, can be measured while an individual is walking. Longer double support time, can also be correlated to balance [26].

5) *Outstanding Challenges in Gait Analysis*: A challenge facing this realm of research is that aging causes many of these occurrences. Therefore successful research in this field has looked not only for the presence of these indicators, most of which are likely common in an aging population, but also at the extent to which an indicator is present. This aims to separate changes typical of aging from those of dementia onset. However, this continues to be a challenge in this field of study [24], [25]. In addition, since the indicators at this stage are largely behavioral and physiological it is very challenging to identify these in a clinical setting. In addition, the way these indicators present in each patient differs, and, in many cases, a patient may experience only a subset of all the indicators mentioned. Adjusting for individual differences and variation during analysis is another challenge of this field.

C. Sensors for Collecting Gait Measures

As discussed above, there are many different metrics we can use to describe gait, including distance walked, speed, disturbances or difficulties, pace, rhythm, variability, stability and width of base. Most research conducted on gait analysis in the clinical setting relies on health care providers observing and manually identifying the presence or prevalence of certain features [29]. We have the technological capabilities to streamline this into a single process. There are three primary classes of technology being used in gait analysis: pressure-based, force-based, and mobile device sensors.

Pressure-based and force-based systems have been developed by numerous groups as a way to monitor gait. These sensors are typically designed as an insole with multiple pressure sensors throughout the insole or a treadmill that can identify how much weight is being exerted in a location at a particular time. From this, center of mass and the elements of pace and stability can be derived. In addition, these systems excel at identifying stability, balance, and the base stance [30]. For example, such a system could determine which part of the foot bears most of the weight of the individual when flat footed or if when they are motionless, there is a continuous shift to where their center of mass is, indicating issues maintaining a balanced stance.

The other type of device that is showing great potential in the area of gait detection is mobile device sensors. Mobile devices include a multitude of sensors, including an

accelerometer, digital compass, gyroscope, proximity sensor, ambient light sensor, GPS, microphone, and camera [31]. Using these sensors, a device can be configured to track elements of gait such as stride length, base width, pace, and variability, as well as distance travelled and speed travelled [32]. In addition, applications can use activity recognition algorithms on the device to log data only when the desired activity is present, which improves the specificity of the data being collected.

As can be seen, there is much overlap between the capabilities of pressure-based insole sensors and mobile device sensors. However, a few aspects set them apart. Pressure-based systems lend themselves to tracking balance and stability better than mobile devices, while mobile devices benefit from increased variability of sensors and computing power that makes them capable of more powerful aggregations, measuring speed, distance, and classifying motion.

D. Machine Learning

The problem of diagnosing vascular dementia early in onset is a challenging one. For one, not all symptom patterns or indicators will be present in an individual patient, but rather a subset of these. The ideal system would have to be flexible enough to account for the variability in disease onset, while also not being so flexible as to create a multitude of false-positive results. Another challenge is that behavioral and physiological patterns differ between individuals. To account for this, the system must learn to confirm the individuality of measures to the known pattern, to create personalized and actionable knowledge. There are there are numerous learning models to explore and within each model, multiple attributes that affect the effectiveness of the model. When choosing models to test, we are looking for a two-part classifier. However, the problem is complicated by the fact that not all vascular dementia onsets in the same way. For this reason, the way attributes interrelate will be an important focus of an effective model.

III. METHODOLOGY

Our study focuses on developing a framework for identification and prediction of subcortical vascular dementia. The framework involves analyzing the gait and executive functioning using machine learning. This methodology involves two parts. Part one focuses on gait analysis while part two focuses on executive functioning evaluation as seen in Fig. 4.

Phase 1: Gait Analysis: Our framework proposes force plate gait measurement system that utilizes Piezoelectric measurement technology for accurate force and moment data gathering. Measuring changes in gait pattern or alterations in the center of mass enables tracking the changes in gait after a significant cardiovascular event. Gait analysis can be traced back to the early eighties where the analysis was done to in the assessment of cerebral palsy, neuromuscular disorders and other orthopedic evaluations [33].

Gait, a means of locomotion, is the style or manner of walking that can be measured by two domains of measurement systems: i) Spatiotemporal system-where the

step length, step width, speed and stride frequency are taken into consideration; and ii) Kinematic -where the angles of joints and rotation of knee, hip and ankle are taken into consideration. Some of the parameters associated with gait that are valuable in the prediction of VD are velocity, center of gravity, stride frequency, step length and frequency, and symmetry of limb movement, defined and depicted in Table II and Fig. 1, respectively [7], [34]. In particular, VD is predictive by the following specific characteristics of gait as seen in Table I [34]. The technique for calculating these attributes of motion can be seen in in Fig. 2 and 3.

TABLE I. GAIT PARAMETERS IN VASCULAR DEMENTIA

#	Parameter	Variation in VD
1	Velocity	Slower
2	Stride length	Shorter
3	Posture/Balance	Instability
4	Start	Hesitation/Freezing
5	Stride Frequency	Decreased
6	Gait lines variability	Increased
7	Tandem gait	Increased
8	Arm swing	Decreased

TABLE II. DEFINITION OF QUANTITATIVE GAIT PARAMETERS. ALL QUANTITATIVE PARAMETERS DESCRIBED BELOW ARE AUTOMATICALLY CALCULATED AS THE MEAN OF TWO TRIALS BY THE GAIT SOFTWARE [2], [26], [34]

Variable	Description
Velocity	Distance covered over a period of time
Stride length	Distance between heel points of two steps of the same foot
Stride Frequency	Steps per unit of time
Balance/Posture	Center of gravity variation

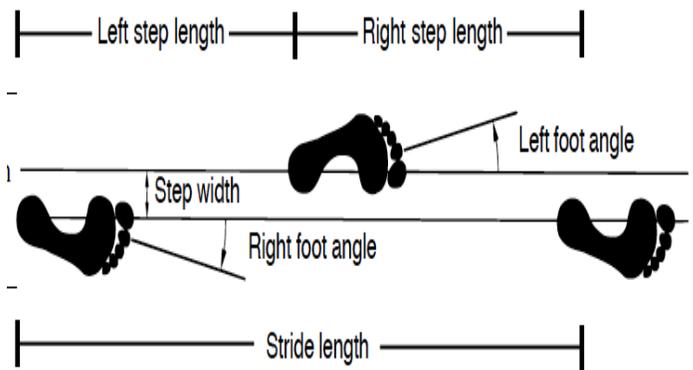


Fig. 1. Step and Stride in Gait: Adapted from [2].

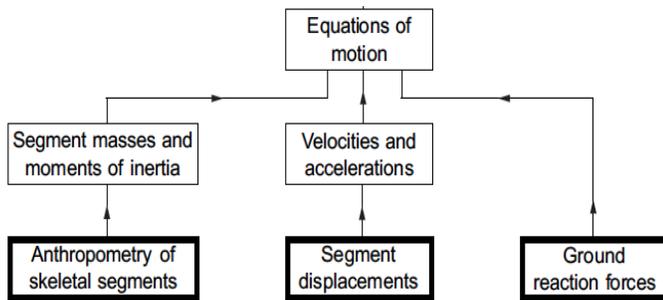


Fig. 2. Gait Parameters and Motion in Words: Adopted from [2].

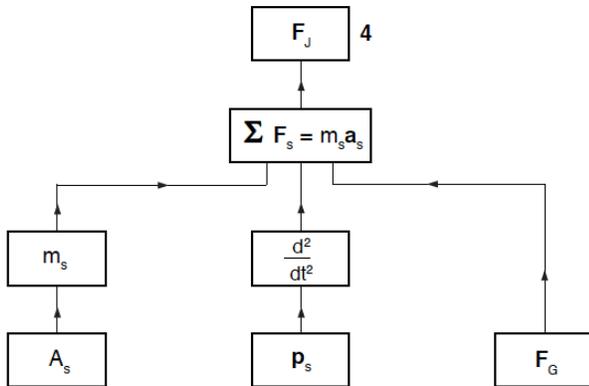


Fig. 3. Formulas for Gait Parameters and Motion: Adopted from [2].

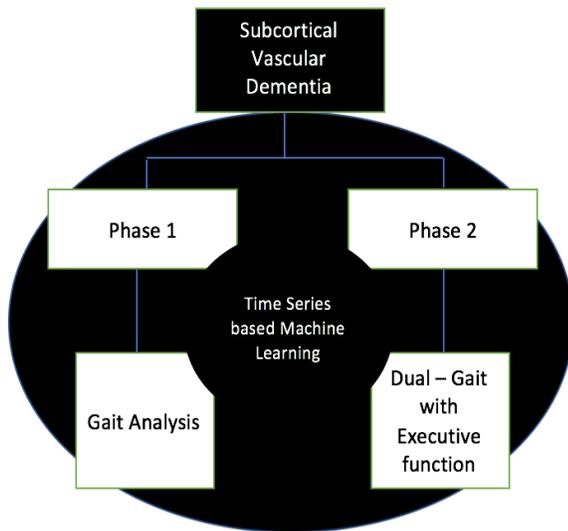


Fig. 4. Framework Combining Gait and Executive Function Analysis Applying Time Series based Machine Learning.

Phase 2: Executive function Analysis: Dual task performance in gait has been recognized as a marker for VD [35], [37]. In the second part of the protocol, the framework performs executive function analysis by having the subject walk and recognizes verbs and walk and finger tap. The evaluation will be performed in two steps:

Step 1: The subject will walk and finger tap

Step 2: The subject will walk and recognize verbs from pictures that are projected on the wall facing the walk.

The Stride length and gait speed will be tracked during this phase, as executive function disorders were recognized as a reduction in gait speed and stride length [33]. One mechanism for examining executive function is by verb recognition or by dual function performance such as finger tapping while walking or verb recognition while walking [33], [35]-[37].

IV. DISCUSSION

A combination of alterations in gait and executive functioning has been identified as a 5 year predictor for VD [8], [9], [16], [38]. Our proposal addresses subcortical vascular dementia that affects gait and executive functioning in affected individuals [17]. Although vascular dementia is marked by its impact on problem solving, execution of complex commands, motor skills, and memory, these domains are not the first to be noticed. VD first impacts gait and executive functioning, and hence we are proposing a framework that uses machine learning to identify alterations in gait and executive functioning [17]. According to the National Institute of Health and Care Excellence (NICE) guidelines the diagnosis of vascular dementia is performed by tools that are based on memory impairment but the impairment is observed only in later stages of VD. Hence we are proposing a system that would diagnose VD earlier than the initiation of the impairment so pharmacological treatments can be utilized to help postpone cognitive impairment. Tests such as minimal state examination (MMSE) do not test for executive functioning impairment, which renders them insignificant tools for Vascular dementia identification. Although Montreal Cognitive Assessment (MOCA) does test for executive functioning impairment and can be used for VD, it relies on factors that can be identified only in later stages of dementia hence reducing its value in treatment of memory impairment. There is a need for tools that can identify and predict VD before its markers are evident. These tools will help clinicians treat individuals with pharmacological remedies that can help minimize cognition impairments as a result of VD.

V. CONCLUSION

The uniqueness of our approach is the application of machine learning along with a combination of gait and executive function analysis. Several research studies have identified gait as an important marker in early identification of vascular dementia [8], [9], [16], [38], while some other researchers have identified executive function as an early marker for subcortical vascular dementia [33], [35]-[37]. We are proposing a framework that combines gait and executive function as a hybrid early predictor of subcortical vascular dementia. The parameters for gait such as reduction in velocity, stride frequency, stride length, and dual cognitive functioning such as finger tapping while walking and verb recognition while walking, with time series based machine learning analysis.

Future work should be focused on implementing a machine learning technique that combines gait and executive functioning metrics utilizing this framework. In addition, research into incorporating other early indicators of subcortical vascular dementia into the framework should be pursued.

REFERENCES

- [1] Plassman, Brenda L., et al. "Prevalence of dementia in the United States: the aging, demographics, and memory study." *Neuroepidemiology* 29.1-2 (2007): 125-132.
- [2] Vaughan, Christopher L., Brian L. Davis, and C. O. Jeremy. "Dynamics of human gait." (1999).
- [3] Román, Gustavo C. "Vascular dementia may be the most common form of dementia in the elderly." *Journal of the neurological sciences* 203 (2002): 7-10.
- [4] Gorelick, Philip B. et al. "Vascular Contributions to Cognitive Impairment and Dementia: A Statement for Healthcare Professionals From the American Heart Association/American Stroke Association." *Stroke; a journal of cerebral circulation* 42.9 (2011): 2672–2713. *PMC*. Web. 5 Apr. 2018.
- [5] Petersen, Ronald C., et al. "Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review) Report of the Quality Standards Subcommittee of the American Academy of Neurology." *Neurology* 56.9 (2001): 1133-1142.
- [6] Bonnici-Mallia, Anne M., Christopher Barbara, and Rahul Rao. "Vascular cognitive impairment and vascular dementia." *InnovAiT* (2018): 1755738018760649.
- [7] Beauchet, Olivier et al. "Guidelines for Assessment of Gait and Reference Values for Spatiotemporal Gait Parameters in Older Adults: The Biomathics and Canadian Gait Consortiums Initiative." *Frontiers in Human Neuroscience* 11 (2017): 353. *PMC*. Web. 13 Apr. 2018.
- [8] Cummings, Jeffrey L. "Vascular subcortical dementias: clinical aspects." *Dementia and Geriatric Cognitive Disorders* 5.3-4 (1994): 177-180.
- [9] Román, Gustavo C., et al. "Vascular dementia Diagnostic criteria for research studies: report of the NINDS-AIREN International Workshop." *Neurology* 43.2 (1993): 250-250.
- [10] Román, Gustavo C., et al. "Subcortical ischaemic vascular dementia." *The Lancet Neurology* 1.7 (2002): 426-436
- [11] P. Chowriappa, S. Dua, and Y. Todorov. "Introduction to machine learning in healthcare informatics". *Machine Learning in Healthcare Informatics*. Springer, 2014, pp. 1–23.
- [12] G. Eysenbach. "Recent advances: Consumer health informatics". *BMJ: British Medical Journal* 320.7251 (2000), p. 171.
- [13] D. L. Sackett. "Evidence-based medicine". *Seminars in perinatology*. Vol. 21. 1. Elsevier. 1997, pp. 3–5.
- [14] E. O. Lillie, B. Patay, J. Diamant, B. Issell, E. J. Topol, and N. J. Schork. "The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?" *Personalized medicine* 8.2 (2011), pp. 161–173.
- [15] M. Panahiazar, V. Taslimitehrani, A. Jadhav, and J. Pathak. "Empowering personalized medicine with big data and semantic web technology: promises, challenges, and use cases". *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE. 2014, pp. 790–795
- [16] Verghese J, Derby C, Lipton R. High risk neurological gaits and vascular dementia. *Neurology*. 2006;66(suppl 2):A57A58.
- [17] V. J. A. Verlinden, J. N. van der Geest, R. F. A. G. de Bruijn, A. Hofman, P. J. Koudstaal, and M. A. Ikram. "Trajectories of decline in cognition and daily functioning in preclinical dementia". *Alzheimer's & Dementia* 12.2 (Feb. 1, 2016), pp. 144–153. issn: 1552-5260. doi: 10.1016/j.jalz.2015.
- [18] R. A. Sperling et al. "Toward defining the preclinical stages of Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". *Alzheimer's & Dementia* 7.3 (May 1, 2011), pp. 280–292. issn: 1552-5260. doi: 10.1016/j.jalz.2011.03.003.
- [19] Berger, A-K., et al. "The occurrence of depressive symptoms in the preclinical phase of AD A population-based study." *Neurology* 53.9 (1999): 1998-1998.
- [20] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, J. L. Cummings, M. deLeon, H. Feldman, M. Ganguli, H. Hampel, P. Scheltens, M. C. Tierney, P. Whitehouse, and B. Winblad. "Mild cognitive impairment". *The Lancet* 367.9518 (Apr. 15, 2006), pp. 1262–1270. issn: 0140-6736. doi: 10.1016/S0140-6736(06)68542-5.
- [21] B. C. Stephan, F. E. Matthews, K.-T. Khaw, C. Dufouil, and C. Brayne. "Beyond mild cognitive impairment: vascular cognitive impairment, no dementia (VCIND)". *Alzheimer's Research & Therapy* 1.1 (July 9, 2009), p. 4. issn: 1758-9193. doi: 10.1186/alzrt4.
- [22] K. Rockwood, C. Macknight, C. Wentzel, S. Black, R. Bouchard, S. Gauthier, H. Feldman, D. Hogan, A. Kertesz, and P. Montgomery. "The diagnosis of mixed dementia in the Consortium for the Investigation of Vascular Impairment of Cognition (CIVIC)". *Annals of the New York Academy of Sciences* 903.1 (2000), pp. 522–528.
- [23] D. S. Knopman. "The initial recognition and diagnosis of dementia". *The American journal of medicine* 104.4 (1998), 2S–12.
- [24] Hausdorff, Jeffrey M., and Aron S. Buchman. "What links gait speed and MCI with dementia? A fresh look at the association between motor and cognitive function." (2013): 409-411.
- [25] J. Verghese, C. Wang, R. B. Lipton, R. Holtzer, and X. Xue. "Quantitative gait dysfunction and risk of cognitive decline and dementia". *Journal of Neurology, Neurosurgery & Psychiatry* 78.9 (2007), pp. 929–935.
- [26] J. Verghese, R. B. Lipton, C. B. Hall, G. Kuslansky, M. J. Katz, and H. Buschke. "Abnormality of gait as a predictor of non-Alzheimer's dementia". *New England Journal of Medicine* 347.22 (2002), pp. 1761–176.
- [27] M. Heydarzadeh, C.-T. Tan, M. Nourani, and S. Ostadabbas. "Gait variability assessment in neuro-degenerative patients by measuring complexity of independent sources". *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE. 2017, pp. 3186–3189.
- [28] T. Imker and C. J. Lamoth. "Gait and cognition: the relationship between gait stability and variability with executive function in persons with and without dementia". *Gait & posture* 35.1 (2012), pp. 126–130.
- [29] K. Kong and M. Tomizuka. "A Gait Monitoring System Based on Air Pressure Sensors Embedded in a Shoe". *IEEE/ASME Transactions on Mechatronics* 14.3 (June 2009), pp. 358–370. issn: 1083-4435. doi: 10.1109/TMECH.2008.2008803
- [30] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. "A survey of mobile phone sensing". *IEEE Communications magazine* 48.9 (2010).
- [31] J. Juen, Q. Cheng, V. Prieto-Centurion, J. A. Krishnan, and B. Schatz. "Health monitors for chronic disease by gait analysis with mobile phones". *Telemedicine and e-Health* 20.11 (2014), pp. 1035–1041.
- [32] Davis III, Roy B., et al. "A gait analysis data collection and reduction technique." *Human movement science* 10.5 (1991): 575-587.
- [33] Morgan, Debra, et al. "The potential of gait analysis to contribute to differential diagnosis of early stage dementia: Current research and future directions." *Canadian Journal on Aging/La Revue canadienne du vieillissement* 26.1 (2007): 19-32.
- [34] Beauchet, Olivier et al. "Gait Analysis in Demented Subjects: Interests and Perspectives." *Neuropsychiatric Disease and Treatment* 4.1 (2008): 155–160. Print.
- [35] Piatt, Andrea L., et al. "Action (verb naming) fluency as an executive function measure: convergent and divergent evidence of validity." *Neuropsychologia* 37.13 (1999): 1499-1503.
- [36] Salthouse, Timothy A. "Relations between cognitive abilities and measures of executive functioning." *Neuropsychology* 19.4 (2005): 532.
- [37] Foster, Jordana Bieze. "Neurologic Gait Predicts Vascular Dementia." *Applied Neurology*, 2006, p. 23.
- [38] Allali, Gilles, Marian Van Der Meulen, and Frédéric Assal. "Gait and cognition: The impact of executive function." *Swiss Archives of Neurology and Psychiatry* 161.6 (2010): 195-199.

Recognition of Ironic Sentences in Twitter using Attention-Based LSTM

Andrianarisoa Tojo Martini, Makhmudov Farrukh, Hongwei Ge

Department of Computer Science & Technology
Dalian University of Technology
Dalian, P. R. China

Abstract—Analyzing written language is an interesting topic that has been studied by many disciplines. Recently, due to the explosive growth of Internet, social media has become an attractive source of searching and getting information for research purposes on written communication. It is true that different words in a sentence serve different purposes of conveying the meaning while they are of different significance. Therefore, this paper is going to employ the attention mechanism to find out the relative contribution or significance of every word in the sentence. In this work, we address the problem of detecting whether a tweet is ironic or not by using Attention-Based Long Short-Term Memory Network. The results show that the proposed method achieves competitive performance on average recall and F1 score compared to the state-of-the-art results.

Keywords—Irony detection; attention; attention mechanism; sentiment analysis; long-short-term memory

I. INTRODUCTION

Nowadays, the Web has become an indispensable source of searching and gaining information because of the quantity and diversity of textual content containing opinions expressed by internet users. Blogs, comments, forums, social networks, reactions or opinions are more and more centralized by search engines. The prodigious measure of data streaming from online social networking and micro-blogging platforms like Twitter, is increasingly attracting the many researchers in the area of sentiment analysis. From these social medias, the automatic detection of irony is, therefore, important for the development of sentiment analysis research, but at the same time it is also an interesting challenge from a cognitive point of view and can help to shed some lights on how human beings use irony as a communicative tool.

Sarcasm and irony are very similar. Generally speaking, irony is employed to convey the opposite meaning of the actual things you say, but its purpose isn't to harm the other person unlike sarcasm which is employed to hurt the other person. According to the Gricean tradition [1], the function of irony is to effectively communicate the opposite of the interpretation of the utterance. However, determining whether a text is ironic or not is a difficult task since the differences between ironic and non-ironic texts are usually extremely delicate. For example, one tweet wrote that "Love this weather #not" is ironic, but a similar tweet which wrote "Hate this weather #not happy" is considered as non-ironic.

In this paper, we introduce the deep learning representation in ironic tweets detection tasks by merging the attention mechanism with the LSTM layers and compare it with the state-of-the-art feature engineering approaches, as we know that state-of-the-art irony and sarcasm detection systems often only rely on deep and sequential neural networks [2] [3].

The Section 2 of this paper is a survey of the related work while Section 3 presents the proposed work by explaining the architecture and the methods used. In Sections 4 and 5 the experiment setup and the results are being respectively discussed. Finally, Section 6 presets the conclusion part.

II. RELATED WORK

Identifying the ironic texts can help to understand the social web better and there are many related applications like sentiment analysis. Irony detecting techniques are important to enhance the performance of sentiment analysis. In [4], authors used the LIBSVM to perform the inductive learning for the training dataset perhaps in accordance with the recent work which has explored the use of Support Vector Machines for text classification with more precise results compared to the other classification techniques.

In [5], authors use Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Attentive RNN in irony detection tasks, and compare the results with the state-of-the-art feature engineering approaches. The first one is Convolutional Neural Network (CNN), which is introduced by [6], and used as a sentence modeling technique in Natural Language Processing (NLP) [7] by using word embedding. Their CNN is applied with one-directional convolutions over the embedded word vectors with multiple filters in various sizes. After applying one-max-pooling over all the outputs filters, the scalars are concatenated together as the encoded vector. The second model is Recurrent Neural Network (RNN), which has been created for the use of sequential data. The Neural Network generates an output vector which considers not only the current input, but also the previous result. The last output vector is taken as the encoded vector.

In [8], the authors made some improvements on previous work [9] by adding some features as well as the word graph similarity score. Each tweet is represented as directed unweighted word graph and the edge between each word is created based on the vicinity window size. Each class in the dataset is represented as directed unweighted graphs. Then a vector is produced after comparing each class graph. And this

vector is used as features by machine learning algorithm. The graph is constructed based on a class assignment and then they measure the similarity of a tweet with each class graph.

Some works have also been carried out for detecting satire in English text, for example [10]. Firstly, authors introduce approach to binary classification of satire in English text. Secondly, they propose a list of generalized linguistic features which provide good results on different types of satire corpora. Furthermore, they make available a standard satire corpus which was retrieved from twitter (with user generated tags such as #satire, #satirical). But developed system might not perform very well on time-based satirical posts on social media platforms.

III. PROPOSED APPROACH

A. Self-Attention Mechanism

First of all, since the research is concentrated on the attention mechanism, we have to discuss about the Self-Attention Mechanism. Recurrent Neural Networks (RNNs) output their hidden state h_i as they process a sequence and that hidden state holds a summary of the information in the sequence. We used a self-attention mechanism [11] to amplify the contribution of important words in the final representation.

After using the attention mechanism, we compute r as combination of all h_i (Fig. 1). The weights a_i have been learned by the network and the magnitude of those weights learned signifies the importance of each hidden state in the final representation.

The hidden state at the last time-step is used as the representation of the input. In long sequences case, the Recurrent Neural Network might not be able to hold all the important information in its final hidden state. In order to amplify the contribution of important elements in the final representation, an attention mechanism has been used.

$$r = \sum_{i=1}^{N=1} a_i h_i \tag{1}$$

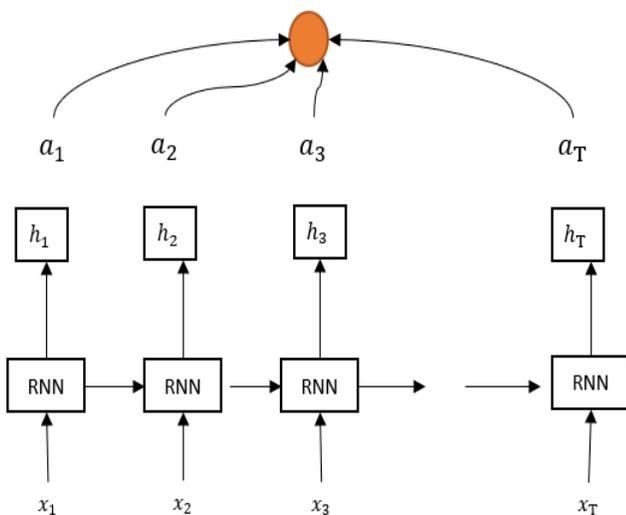


Fig. 1. Attention RNN.

B. Preprocessing

We've used a text processing tool called Ekphrasis presented by [12], which can perform tokenization, word normalization, word segmentation (for splitting hashtags) and spelling correction, using word statistics from two big corpora namely English Wikipedia and Twitter.

1) Tokenization: Tokenization is the initial preprocessing stage which makes it the foundation for the latter stages. Therefore, it will certainly make an effect of the feature's quality studied by the network. Tokenization in Twitter is full of challenges for that various usage of vocabulary and expressions are here and there. Of course, some of the challenges came from the dilemma of projecting the whole expression or simply taking its tokens. To rise to this challenge, Ekphrasis recognized the markup, emoticons, emojis, dates, acronyms, censored words and words with emphasis.

2) Normalization: Apart from the method of tokenization, we also make some adjustment on certain selected tokens, such as spelling correction, words normalization and sedimentation. Furthermore, we also figure out what kinds of tokens should be omitted, normalized and surrounded together with those that should be replaced with special tags such as URLs, emails and @user.

C. Attention-based LSTM Model Description

The framework of our attention-based LSTM network is illustrated in (Fig. 2). Next, we will introduce each layer in our model from bottom to top in detail.

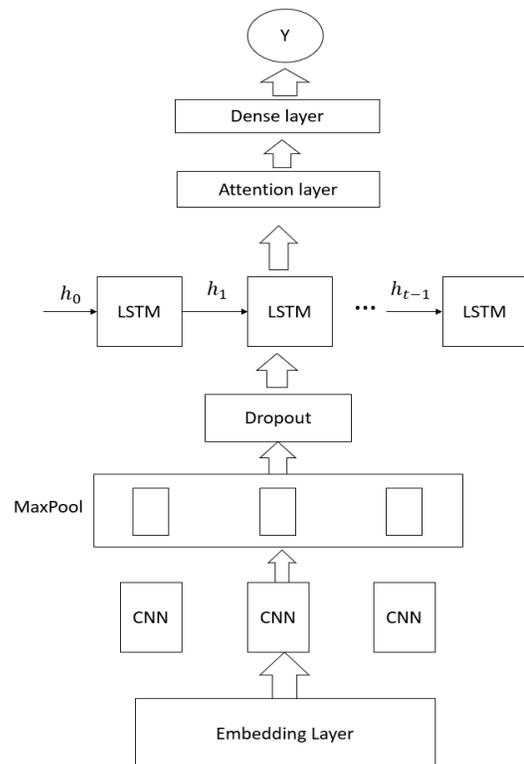


Fig. 2. Architecture of LSTM with Attention Mechanism.

3) Embedding Layer: This process happens just right after the pre-processing. Word embedding techniques aim to use continuous low-dimension vectors representing the features of the words [13], which tweets are transformed into a sequence of words $S = (s_1, \dots, s_N)$, $S \in \mathbb{R}^{N \times d}$, where N is the number of a tweet, and d denotes the dimension of a word vector [14]. We use Word2Vec [13] as the vector representation of the words in tweets.

4) Convolutional and Max-Pooling Layers: After getting the pre-trained word vectors “word2vec” from the word embedding Layers, we train a convolutional neural network, followed by a max-pooling layer. The goal of convolution is to extract the input feature, and pooling is to subsample the output of the convolution matrix. The regular way to do pooling is by applying a max operation to the result of each filter. There are two reasons to use a max-pooling layer in our research. First, by doing elimination of any non-maximal values, it reduces computation for upper layers. Second, the max-pooling layer can extract the local dependency within different regions to keep the most salient information.

5) LSTM Layer: The next layer in our model is LSTM layer. LSTM is kind of RNN which has been introduced firstly by [15]. For LSTM, Cell state (c_t) are connected to three gates which are forget gate (f_t), input gate (i_t) and output gate (o_t) respectively. Fig. 3 illustrates the architecture of a standard LSTM.

More formally, each cell in LSTM can be computed as follows:

$$X = \begin{bmatrix} h_{t-1} \\ x^t \end{bmatrix} \quad (2)$$

$$f_t = \theta(W_f \cdot X + b_f) \quad (3)$$

$$i_t = \theta(W_i \cdot X + b_i) \quad (4)$$

$$\tilde{C}_t = \tanh(W_c \cdot X + b_c) \quad (5)$$

$$C_t = f_t * C^{t-1} + i_t * \tilde{C}_t \quad (6)$$

$$o_t = \theta(W_o X + b_o) \quad (7)$$

$$h_t = o_t * \tanh(C_t) \quad (8)$$

Where $W_f, W_i, W_o, W_c \in \mathbb{R}^{d \times 2d}$ are the weighted matrices and $b_f, b_i, b_c, b_o \in \mathbb{R}^d$ are biases of LSTM to be learned during training, parameterizing the transformations of the input, forget and output gates respectively. θ is the sigmoid function and $*$ stands for element-wise multiplication, x^t includes the inputs of LSTM cell unit.

This layer is used to capture long-range contextual information from tweets. At time step i , a hidden state h_i is generated which contains both previous and future context information. Since different words and phrases serve different purposes to irony detection, we propose to design an attention layer after the LSTM layer to help our model focus on important words and contexts.

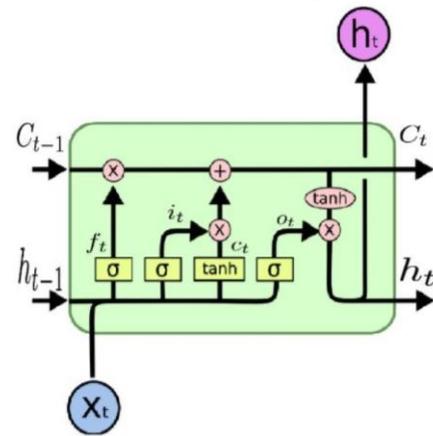


Fig. 3. Architecture of Cell in LSTM.

6) Attention Layer: The input of the attention layer is the hidden state vector h_i at each time step. The attention weight m_i for this time step can be computed as:

$$m_i = \tanh(h_i) \quad (9)$$

$$\hat{\alpha}_i = w^T m_i + b \quad (10)$$

$$\alpha_i = \frac{\exp(\hat{\alpha}_i)}{\sum_j \exp(\hat{\alpha}_i)} \quad (11)$$

Where w and b are the parameters of the attention layer. The output of attention layer at the i_{th} time step is formulated as follows:

$$r_i = \alpha h_i \quad (12)$$

IV. EXPERIMENTAL SETUP

First of all, let's talk about the datasets. The dataset used consists of 355k English tweets (43k ironic and 312k in literal sentiment sense, we named it dataset1. Another dataset collected by Ghosh [2] contains 18k sarcastic tweets (which can be used on irony) and 21k regular tweets. In order to collect the most data for dataset1, we used the Twitter API ([https:// dev.twitter.com/](https://dev.twitter.com/)) to stream tweets from Twitter by using hashtags #irony, #sarcasm and #not as key word. And the data was cleaned by using the preprocessing method from the section 3 (which means that ironic hashtags, such as #not, #sarcasm, #irony, in the dataset have been removed), it was labeled 1 for ironic texts and 0 for normal.

As for the implementation, our model is implemented in Keras library. We conducted the experiment with different values for the LSTM hidden state size and for the dropout probability, obtaining best results for a dropout probability of 0.5 and 128 units for the hidden vector. The table below (Table I) shows the repartition of the collected dataset, we trained 80% of the provided data as training set and 20% as test set. Since the data is kind of voluminous, we only use the number of epochs as 3. Cross entropy and Adam are used as the loss function and optimization algorithm of the output layer.

TABLE I. COUNTS AND PERCENTAGES OF IRONIC AND NON-IRONIC OF THE TWEETS COLLECTED AND TEST-TRAIN SET

	Non-Ironic	Ironic	Total
Training set	249800 (88%)	34382 (12%)	284182
Test set	62501 (88%)	8545 (12%)	71046
Collected data	312193 (88%)	43035 (12%)	355228

V. FINAL RESULT AND DISCUSSION

A. Results

Tables II and III show the results of the experiments after using both LSTM approach and Attention Based approach, and compare them to the state models presented by [2]. We only report the average Precision (Avg.Prec), Recall (Avg.Rec), and F1 scores (Avg.F1).

Table II below presents a comparison of the results trained on the collected dataset (dataset1), we observe that our model with Attention based LSTM almost outperforms every model than other models, except the model which is a combination of CNN, LSTM, and DNN introduced by [2], it outperforms our model at the precision by 0.4% margin but they both got the same results on the F1 score. As for the proposed model with just LSTM, it performs the lowest performance in every evaluation.

As for Table III, we show that the performance of our system can outperform some of the baseline methods on the Ghosh dataset [2] but got outperformed by the CNN, LSTM, and DNN model.

TABLE II. COMPARISON OF OUR METHOD TO BASELINE USING DATASET1

Model		Avg. Prec	Avg. Rec	Avg. F1
Our model	Attention based LSTM	0.836	0.883	0.859
	LSTM	0.703	0.805	0.751
Ghosh	CNN + LSTM + DNN (with dropout)	0.84	0.876	0.857
	LSTM+ LSTM	0.734	0.842	0.784
	CNN+CNN	0.716	0.804	0.758

TABLE III. COMPARISON OF OUR METHOD TO BASELINE USING GHOSH DATASET

Model		Avg. Prec	Avg. Rec	Avg. F1
Ghosh	CNN + LSTM + DNN (with dropout)	0.899	0.91	0.904
	LSTM+ LSTM	0.854	0.871	0.862
	CNN+CNN	0.856	0.879	0.868
Our model	LSTM	0.777	0.859	0.816
	Attention based LSTM	0.875	0.894	0.884

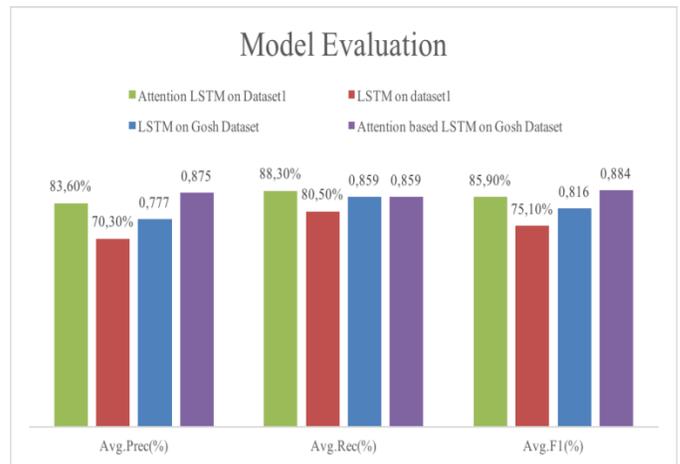


Fig. 4. Attention Architecture with LSTM with Attention Mechanism.

Fig. 4 shows that when using the Attention Mechanism on the LSTM layer, the model performs better than the one that doesn't use it. The Attention based Model makes an improvement on the Precision by more than 9%, around 3 to 8% on Recall and more than 8% on F1 score.

B. Discussion

1) Attention visualization: In the following figure (Fig. 5), we are going to get a closer look at the degree showing how much attention mechanism will better the performance of irony detection.

According to the given figure, there are some certain usage of language such as apparent emotional words, old topics, emojis, punctuation, numerals and sometimes slang and ungrammatical expressions attaining much more focus in the internet which makes it the biggest factor in case of the contribution to irony detection. The network is going to study the significance of certain words, it targets at finding out what factors will make a difference when it comes to the final ironical decision. As shown in the figure, the reddish color is used to highlight attention weights and the color gradients are there to make a distinction between the heavy weights of attention and the light one.

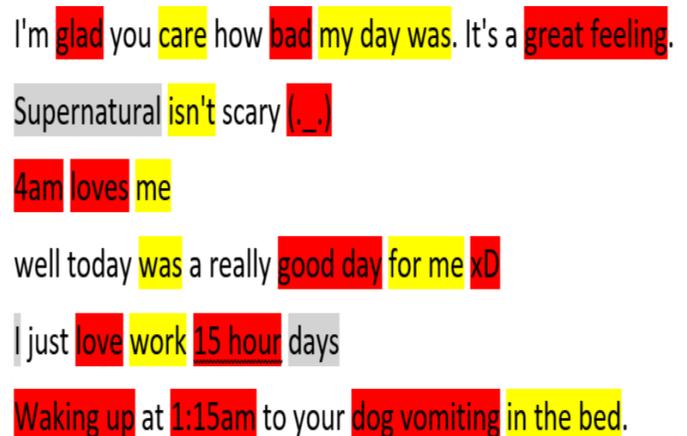


Fig. 5. Attention Visualization.

VI. CONCLUSION

In this paper, we proposed a Long Short-Term Memory (LSTM) with attention mechanism model to detect English ironic sentences from Twitter. The proposed model got competitive result compared to the state-of-the-art models without using further feature engineering. The results showed that our model performs better on the collected dataset, especially on the recall and f1 score. On the Ghosh [2] dataset, our Attention-Based model outperformed the CNN and LSTM model proposed by [2] but couldn't outperform the model with a combination of CNN, LSTM, and DNN. Finally, in the discussion part we show that the attention vectors generated by our attention layer can capture specific words which are very useful to decide for the training, it can decide whether the tweet selected is ironic or not. In a future work, we would like to explore how to make full usages of the attention mechanism on text sentiment analysis.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China No. 61471084, No. U1608253, and the open program of State Key Laboratory of Software Architecture No. SKLSA2016B-02.

REFERENCES

- [1] S. Chapman, "Logic and Conversation," in Paul Grice, *Philosopher and Linguist*, London, Palgrave Macmillan UK, 2005, pp. 85-113.
- [2] A. Ghosh and D. T. Veale, "Fracking Sarcasm using Neural Network," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, San, 2016.
- [3] M. Zhang, Y. Zhang and G. Fu, "Tweet Sarcasm Detection Using Deep Neural Network," in *COLING*, 2016.
- [4] T. Ahmad, H. Akhtar, A. Chopra and M. W. Akhtar, "Satire Detection from Web Documents Using Machine Learning Methods," 2014 International Conference on Soft Computing and Machine Intelligence, pp. 102-105, 2014.
- [5] Y.-H. Huang, H.-H. Huang and H.-H. Chen, "Irony Detection with Attentive Recurrent Neural Networks," in *ECIR*, 2017.
- [6] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278-2324, 11 1998.
- [7] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.
- [8] U. Ahmed, L. Zafar, F. Qayyum and M. Arshad Islam, "Irony Detector at SemEval-2018 Task 3: Irony Detection in English Tweets using Word Graph," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, New, 2018.
- [9] G. Giannakopoulos, V. Karkaletsis, G. Vouros and P. Stamatopoulos, "Summarization System Evaluation Revisited: N-gram Graphs," *ACM Trans. Speech Lang. Process.*, vol. 5, pp. 5:1--5:39, 10 2008.
- [10] A. N. Reganti, T. Maheshwari, U. Kumar, A. Das and R. Bajpai, "Modeling Satire in English Text for Automatic Detection," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.
- [11] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *CoRR*, vol. abs/1409.0473, 2014.
- [12] C. Baziotis, N. Pelekis and C. Doulkeridis, "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, 2017.
- [13] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *CoRR*, vol. abs/1301.3781, 2013.
- [14] Y. Zhang, J. Wang and X. Zhang, "YNU-HPCC at SemEval-2018 Task 1: BiLSTM with Attention based Sentiment Analysis for Affect in Tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, New, 2018.
- [15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, pp. 1735-1780, 1997.

The Role of Camera Convergence in Stereoscopic Video See-through Augmented Reality Displays

Fabrizio Cutolo

University of Pisa

Dept. of Information Engineering & EndoCAS Center
Via Caruso 16, 56122, Pisa

Vincenzo Ferrari

University of Pisa

Dept. of Information Engineering & EndoCAS Center
Via Caruso 16, 56122, Pisa

Abstract—In the realm of wearable augmented reality (AR) systems, stereoscopic video see-through displays raise issues related to the user’s perception of the three-dimensional space. This paper seeks to put forward few considerations regarding the perceptual artefacts common to standard stereoscopic video see-through displays with fixed camera convergence. Among the possible perceptual artefacts, the most significant one relates to diplopia arising from reduced stereo overlaps and too large screen disparities. Two state-of-the-art solutions are reviewed. The first one suggests a dynamic change, via software, of the virtual camera convergence, whereas the second one suggests a matched hardware/software solution based on a series of predefined focus/vergence configurations. Potentialities and limits of both the solutions are outlined so as to provide the AR community, a yardstick for developing new stereoscopic video see-through systems suitable for different working distances.

Keywords—Augmented reality and visualization; stereoscopic display; stereo overlap; video see-through

I. INTRODUCTION

Human eyes are placed frontally about 6-7 cm apart (interpupillary distance = 6-7 cm) so they both perceive the same scene but from slightly different viewpoints (i.e. with an horizontal parallax). In other words, through the crystalline lenses, the two retinas receive slightly different views of the same three-dimensional (3D) scene. The positional differences between the two retinal images are defined as binocular or retinal disparities. Specialized neurons (binocular neurons) in the visual cortex of the brain, process those disparities to generate a sort of depth map of the observed scene. We commonly refer to this mechanism as stereopsis from the Greek words $\sigma\tau\epsilon\rho\epsilon\omicron$ (stereo-meaning solid and $\omicron\psi\upsilon\varsigma$ (opsis meaning appearance, sight, and we define this depth cue as binocular parallax. The goal of stereoscopic 3D displays is hence to create an illusion of depth perception by providing consistent binocular disparity information in the recorded images delivered to the left and right eyes [1].

Depth cueing through stereoscopy is an essential feature of head-mounted displays (HMDs) for augmented reality (AR). Most of the AR HMDs fall into two categories according to the see-through paradigm they implement: video see-through HMDs and optical see-through HMDs. Typically, in optical see-through systems, the user’s direct view of the real world is augmented with the projection of virtual information on a beam combiner and then into the user’s line of sight [2]. Differently, in video see-through systems the virtual content is

merged with camera images captured by a stereo camera rig rigidly fixed on the 3D display.

The pixel-wise video mixing technology that underpins the video see-through paradigm can offer high geometric coherence between virtual and real content. Nevertheless, the industrial pioneers, as well as the early adopters of AR technology properly considered the camera-mediated view typical of video see-through devices as drastically affecting the quality of the visual perception and experience of the real world [2], [3].

In stereoscopic video see-through HMDs, double vision (diplopia) for the user may arise if the fixation point, determined by the intersection of the optical axis of the stereo camera pair, leads to reduced stereo overlap between the two images delivered to the eyes through the HMD. This stereo conflict happens because a large portion of the scene is not represented on both images (the left part of the left view is not represented in the right view and vice versa), and therefore the visual cortex is not able to fuse the two views. This perceptual conflict is due to the fixed configuration of the stereo setting and it heavily constraints the working distance on where the stereoscopic video see-through HMD yields a comfortable visual result. Two possible solutions for coping with this limit are here reviewed and brought as example. This work is inspired by the need for assessing the role of camera convergence in stereoscopic video see-through AR displays and establishing a yardstick for designing new solutions that allow users of such systems to work comfortably at close distance ranges.

II. BINOCULAR PARALLAX, CONVERGENCE AND ACCOMMODATION IN HUMAN VISUAL SYSTEM

A. Binocular Parallax, Horopter and Panum’s Area

Binocular parallax is the most effective relative depth cue at close distances, namely in an individual’s personal space or at arm’s length [4], [5]. The equation that links the theoretical depth discrimination threshold (i.e. human depth resolution) ΔZ_h to the angular retinal disparity $\Delta\alpha$ can be trivially derived from geometrical relations (see Fig. 1). In particular, for a given a fixation point, associates the convergence angle of the eyes θ to the absolute depth of the fixation point (Z) and to the interpupillary distance I :

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731974-VOSTARS project www.vostars.eu

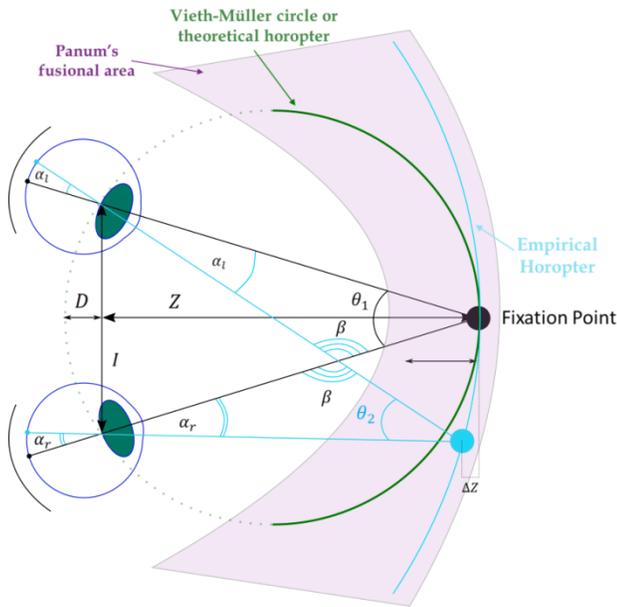


Fig. 1. Binocular horizontal disparity and geometry of the stereo perception in human visual system. Human depth resolution can be approximately expressed as a function of the distance between fixation point and observer Z , and of the interpupillary distance I .

$$\theta = 2 \tan^{-1}(I/2Z) \quad (1)$$

The human depth resolution formula [6] is the result of the derivative $dZ/d\theta$:

$$\Delta Z_h = -\frac{Z^2}{I} \left(1 + \frac{I^2}{4Z^2}\right) \Delta\theta \quad (2)$$

Whose simplified form is:

$$\Delta Z_h \approx -\frac{Z^2 * \Delta\theta}{I} \quad (3)$$

It is worth noting, from trivial geometrical considerations that taking into account the angle α :

$$\begin{aligned} \Delta\theta &= \theta_1 - \theta_2 = (\pi - \beta - \alpha_l) \\ &\quad - (\pi - \beta - \alpha_r) = \alpha_r - \alpha_l \\ &= -\Delta\alpha \end{aligned} \quad (4)$$

Hence,

$$\Delta Z_h \approx \frac{Z^2 * \Delta\alpha}{I} \quad (5)$$

So, and as explained with more details in [7], (5) associates the depth resolution of the human visual system to the retinal angular difference $\Delta\alpha$, to the interpupillary distance I , and to the distance between fixation point and observer Z [8].

When retinal disparities are too high, they produce diplopia that is, itself, a depth cue for the human visual system. The fixation point has 0 retinal disparity, as well as all the corresponding retinal points. The circle formed passing through the fixation point F and the two nodal points of the eyes O_1 and O_2 , is defined Vieth-Müller circle or theoretical horopter (from the Greek words $\delta\rho\omicron\varsigma$ meaning boundary, + $\delta\pi\tau\eta\rho$ meaning observer). Any points belonging to this circle stimulate geometric corresponding points on the retinae of the two eyes, hence they bring 0 disparity exactly as F ($\alpha_l = \alpha_r$ in Fig. 1 for construction).

In reality, the empirical horopter for any observer is less convex than a circle and the Hering-Hillebrand deviation gives a measure of the deviation of the empirical horopter from the Vieth-Müller circle [8], [9]. According to such deviation (referred to as H), the relation that fits the empirical measurements on the real horopter based on the disparity between the retinal angles α_l and α_r is (see Fig. 1):

$$H = \cot \alpha_l - R \cot \alpha_r \quad (6)$$

that leads to [10]:

$$\alpha_r = \tan^{-1} \frac{R \tan \alpha_l}{1 - H \tan \alpha_l} \quad (7)$$

with R = relative magnification between right eye and left eye. Thus the empirical deviation from the theoretical horopter is measured by disparity D between α_l and α_r and it is modeled as follows:

$$D = \alpha_l - \tan^{-1} \frac{R \tan \alpha_l}{1 - H \tan \alpha_l} - D_0 \quad (8)$$

with D_0 encapsulating the effect of the Helmholtz shear. The conditions for the theoretical horopter are: $D_0 = 0$, $H \approx 0$, and $R \approx 1$. In that case, the empirical disparity of the points belonging to the horopter is null. Within a special visual space around the fixation point (except the points belonging to the horopter), the corresponding points on the retina produce disparities whose values are processed by the visual cortex of the brain to estimate depth relations in such area around the fixation point (Fig. 2). These disparities, provided they are sufficiently small, can be either positive, if the points under observation are behind the horopter, or negative if the points are in front of the horopter. This visual space is called Panum's fusional area. In this area, the two stereoscopic retinal images are fused into one. Outside the Panum's area, the objects are perceived as diplopic.

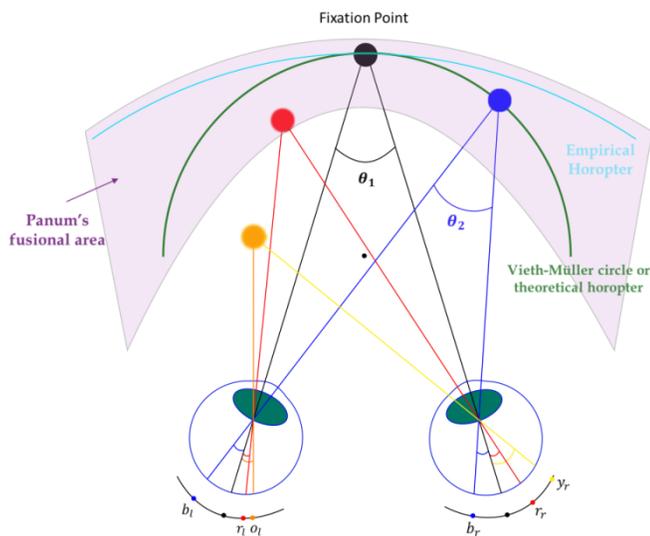


Fig. 2. At a given distance of the fixation point, any point belonging to the horopter is imaged on geometric corresponding retinal points ($\mathbf{b}_l \equiv \mathbf{b}_r$). Within a special visual space around the fixation point (other than the points belonging to the horopter), the projected points on the retina produce disparities that are mapped into depth perception around the fixation point. This visual space surrounding the horopter is called Panum's fusional area. In this area, the two stereoscopic retinal images are fused into one by the brain. Outside the Panum's area, the objects are perceived as diplopic (\mathbf{o}_l cannot be fused with \mathbf{y}_r).

B. Convergence and Accommodation

In human visual system, convergence is that disjunctive movement of the eyes that allows the minimization of the targeted visual information projected on the two retinas [11]. The coordinated action of the extra-ocular muscles, when focusing on the same object, causes the mutual rotation of the optical axes of the eyes and, therefore, helps in perceiving depth/distance in combined interaction with accommodation. As well as the accommodation cue, this cue is powerful within the personal space [12].

III. STEREOSCOPIC VIDEO SEE-THROUGH

In stereoscopic video see-through HMDs, the goal is to create an illusion of depth perception by providing consistent binocular disparity information in the images delivered to the left and right eyes by the two displays of the visor.

Any stereoscopic video see-through display comprises two stages whose specifics are to be matched one another in order to provide a consistent illusion of depth perception to the viewer: acquisition stage and viewing stage [13]. The stereo rig anchored to the visor has the task of capturing the real scene (i.e. acquisition task), whereas the two internal displays of the visor have the task of delivering the stereoscopic augmented information to the viewer (i.e. visualization task).

In his work, Kyto has extensively addressed all the factors that influence depth perception through stereoscopic video see-through displays [5], [14]. In particular the authors proposed a useful theoretical comparison between human depth resolution through stereo displays and stereo camera depth resolution [15]. By carefully evaluating the results of

that analysis, we can get an idea of the requirements in terms of disparity accuracy Δd , focal length, camera sensor width S_w , and baseline for the external stereo camera pair that allows the achievement of human-like viewing conditions. Nonetheless, in most applications, a tradeoff is to be sought between accuracy in stereo depth measurements (at least comparable to that of the human visual system), and the quasi-ortho-stereoscopic depth perception through the video see-through HMD [14]; for example a changing of the baseline and/or of the focal length may lead to an improved depth resolution out of stereo triangulation, but at the expense of introducing unwanted perceptual artefacts to the viewer [13], [16].

Particularly in image guided surgery, the quality of an AR experience depends on how well the virtual content is integrated into the real world spatially, photometrically and temporally [17]. In this context, wearable AR systems offer the most ergonomic solution for those medical tasks manually performed under user's direct vision (open surgery, introduction of biopsy needle, palpation, etc.) since they intrinsically provide the user with an egocentric viewpoint of the surgical scene. They contextually integrate the surgeon's visual perception of the real scene with useful AR-based visualization modalities (derived from radiological images). Different embodiments of video see-through HMDs have been proposed in minimally invasive surgery [18], [19], laparoscopic surgery [20]-[22], orthopedic surgery [23], [24], in neurosurgery [25], and in maxillofacial surgery [26], [27]. In assessing the efficacy and reliability of such devices, the understanding of all the physiological and psychological mechanisms that underpin depth perception is of particular importance. In this regard, unreliable modalities of AR visualization can in fact bring cognitive overload and perceptual conflicts causing misinterpretation and hindering clinical decision-making [28].

A comprehensive overview of all the possible perceptual artefacts that arise in the acquisition or in the visualization stage in stereoscopic video systems is presented in [13]. Among all the possible perceptual artefacts, as anticipated, diplopia may arise if the fixation point, determined by the intersection of the optical axes of the stereo camera pair, leads to reduced stereo overlap. In the next subsection, we shall briefly describe two possible solutions for coping with this problem and we shall contextually point out the strengths and weaknesses of both the approaches. It is worth mentioning that both the solutions that we shall review were properly designed for specific medical/surgical applications in which the user is asked to interact with the augmented scene at varying working distances (however at close range), during procedures demanding for high hand-eye coordination. This task-oriented requirement increases the need for stereoscopic video see-through systems that allow sufficient stereo overlap when viewing close objects, although at odds with a desired ortho-stereoscopy [29]. Both solutions in fact feature a non-negligible eye-camera offset, so in rigorous terms, the ortho-stereoscopy of both the systems was not ensured from the very beginning and in contrast with the assertions made by Takagi et al. [30].

IV. VIDEO SEE-THROUGH HMD WITH DYNAMIC VIRTUAL CONVERGENCE

State et al. [16] proposed a software solution with dynamic control of the virtual convergence of the display frustum to allow users to work comfortably at different depths. A narrower selection of the wide-angle cameras imaging frustum is cropped dynamically as a function of a heuristic estimation of the working distance. For each eye, the modified augmented image is then delivered to the corresponding internal monitor of the video see-through HMD. No recalibration of the stereo camera rig is needed since there are no moving parts in the AR system. Nothing is said regarding the camera focuses, so we assume that both were kept fixed for each working distance.

Their solution offers two possible methods for managing the geometry of the display frustum: sheared frustums and rotated frustums. Sheared display frustum does not add unwanted vertical disparity to the stereo images which could bring geometrical artefacts in perceiving depth relations (e.g. keystone depth plane distortion [13], [30]). Unfortunately sheared frustums, especially at close working distances, bring a more pronounced disparity-vergence conflict if compared to the rotated solution. Rotated frustums, albeit introducing vertical disparity between corresponding features in stereo images, especially at the corners, is able to more consistently simulate the physical rotation of the displays and hence to stimulate the user's eyes to converge. This fact, besides reducing the disparity-vergence conflict, provides an additional depth cue.

Another interesting aspect of their work was the real-time and automatic control of the virtual convergence as a function of an estimation of the working distance. The control of the virtual convergence was implemented through three distinct approaches: an image-based method based on the maximization of the mutual information among paired views of the virtual content. A second method, a pixel-wise inspection of the z-buffer of the stereo images is used to provide depth estimation. A third approach, in which a depth estimate is computed by working on geometry data instead than on a pixel-wise mapping of the rendered images. The third method can work on the current frame, before being rendered on the display, and therefore provide an instantaneous managing of the convergence. On the contrary, the first two methods are suitable for predicting convergence for the subsequent frame.

The main drawback of this technique is that it drastically reduces the resolution of the images acquired by the stereo cameras. Further, the idea of dynamically changing the virtual convergence of the stereo camera pair through a real-time estimation of the operative depth, albeit appealing, is prone to possible perceptual conflicts for the user and it may lead to incorrect depth perception or discomfort during use if not properly managed.

V. VIDEO SEE-THROUGH HMD WITH ADJUSTABLE CAMERA VERGENCE

In 2014 Ferrari et al. [31] proposed a matched hardware/software solution that entails the adjustment of the degree of convergence of the stereo camera pair established as a function of the working distance on a per-session basis.

In more details, to restore stereo overlap, and reduce image disparities well within the binocular fusional area, the degree of convergence of the stereo camera pair is made adjustable so to be adapted at different and predefined working distances. In this way the fixation point is moved closer to the observer and the visual disparities between left and right images are reduced (Fig. 3). To implement this idea, an ad hoc version of a previously presented video see-through system [32]-[35] based on a HMZ-T2 Sony HMD, was assembled (Fig. 4). The system comprises two supports equipped with adjuster screws for modifying the stereo camera vergence, and the camera focus can be coherently adjusted with the working distance thanks to motorized mechanisms.

For each set of predefined focus/vergence configuration, the intrinsic and extrinsic camera parameters associated with it are to be determined offline as a result of a one-time calibration routine, with the calibration data stored for subsequent reuse. A two-stage video-based pose estimation algorithm, allows sub-pixel registration accuracy in the augmented scene without requiring additional work to the user (i.e. no further calibrations are required). More specifically:

1) 3D localization through stereo triangulation is correctly performed solely relying on the sets of predetermined intrinsic and extrinsic calibration data associated to the specific stereo camera vergence configuration.

2) The initial estimation of the camera pose, computed in closed-form through a standard SVD-based method [36], may result intrinsically inaccurate, given the uncertainty in the estimation of the stereo camera parameters, but it sets up a good initial guess for the subsequent pose refinement step.

The proposed method was tested on a set of 3 predefined configurations of the stereo camera vergence. For each configuration, the two adjuster screws were set to move the fixation point at 30, 100 and 170 cm. Therefore, an estimation of the corresponding convergence angle was computed by substituting I and Z in (1).

For each configuration and before use, the focus of both cameras was adjusted to focus at the fixation point, hence as a function of a set of expected working distances. Accordingly, three sets of intrinsic and extrinsic camera parameters had to be estimated and stored following three offline calibration routines.

However, also this solution has its own weaknesses. In rigorous terms, the converging of the optical axes of the stereo

cameras alone, without a simultaneous and coupled converging of the optical axes of the displays, goes against a desired ortho-stereoscopy, and therefore it might cause perceptual artifacts [30].

As properly hypothesized by State et al. and Ferrari et al. [31], their experience suggests that the distortion of the perceived 3D space is not too severe to hinder the correct use of the stereoscopic video see-through display; obviously this holds true if we consider the really constrained and task-oriented working distances for which the preset vergence configurations were set. Furthermore, and unlike the solution proposed by State et al., the system has “moving parts” and therefore needs for regular recalibrations to cope with the degradation of the stereo calibration over time.

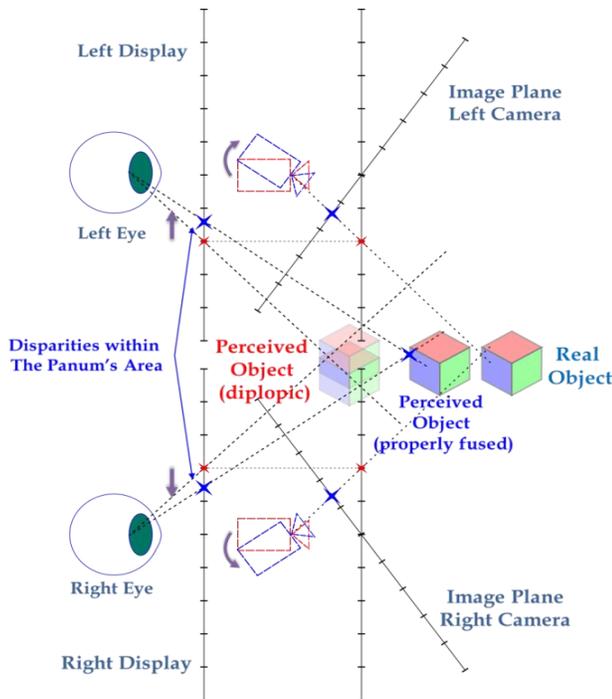


Fig. 3. Adjustment of the degree of convergence of the stereo camera pair to increase stereo overlaps and reduce visual disparities between images on the left and right display of the HMD. The disparity between the blue crosses on the two displays is lower than the disparity between the red crosses.



Fig. 4. HMD prototype embedded with adjuster screws for stereo camera vergence control.

VI. DISCUSSION AND CONCLUSION

The use of stereoscopic video see-through HMDs in case of AR assistance during manually performed tasks (e.g. surgery) in which the user is asked to interact with the scene at close working distances, and during procedures demanding for high hand-eye coordination (e.g. medical/surgical applications), is heavily hindered by the occurrence of diplopia.

Typically, in these systems, cameras and displays are preset at a fixed convergence angle on the basis of assumptions made on the average working distance. Thereby, in these systems, stereo conflicts may arise if the fixation point, determined by the intersection of the optical axis of the stereo camera pair, leads to reduced stereo overlap between the two images delivered to the eyes through the HMD. This occurrence heavily limits the actual distance on where the stereoscopic video see-through HMD can yield a comfortable visual result. In this paper, two possible solutions for coping with this limit were reviewed and brought as example, one purely software and the other matched hardware/software.

The solution suggested by State et al. features a dynamical change of the virtual convergence of the stereo camera pair based on a real-time estimation of the operative depth. This solution does not comprise any moving parts within the HMD, hence is theoretically calibration-free but this is at the expense of a drastic reduction of the resolution of the images acquired by the stereo cameras. Furthermore, having the camera focuses fixed (also to avoid further calibrations) may produce blurred images if the system were used at working distances far from the focus.

The solution by Ferrari et al. entails the physical adjustment of the degree of convergence and of the associated camera focuses of the stereo camera pair, established as a function of the working distance on a per-session basis. In this case, a set of calibration routines (intrinsic and extrinsic) has to be performed before but, in this way, all the camera frustum is viable and the camera images are properly on focus for each working distance established.

Based on our experience and as clearly stated in both the reviewed works, make the augmented scene as stereo-perceivable at close distances is key for those applications in which the user is asked to interact with the augmented scene at close working distances (i.e. in the personal space, at arm's length) and during procedures that demands for high dexterity as a surgical tasks. It is our conviction that this requirement, once achieved, may fully compensate for the increased distortion of the perceived 3D space, due to the dynamic change of the convergence of the cameras (be them virtual or real) [37]. In our opinion, from a functional standpoint, resolving diplopia has a higher priority than dealing with the perceptual artefacts caused to the non-rigorous orthoscopy of the stereoscopic display. Nevertheless, we also believe that the technology should move towards the implementation of parallax-free video see-through HMDs that entails an automatic and coupled management of the display and camera convergence (as a function of a real-time depth estimation algorithm).

REFERENCES

- [1] S. A. Benton, *Selected Papers on Three-dimensional Displays*: SPIE Optical Engineering Press, 2001.
- [2] J. P. Rolland, R. L. Holloway, and H. Fuchs, "A Comparison of Optical and Video See-through Head-Mounted Displays," *Telemanipulator and Telepresence Technologies*, vol. 2351, pp. 293-307, 1994.
- [3] R. T. Azuma, "A survey of augmented reality," *Presence-Teleoperators and Virtual Environments*, vol. 6, pp. 355-385, Aug 1997.
- [4] J. E. Cutting and P. M. Vishton, "Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth."
- [5] M. Kyto, A. Makinen, T. Tossavainen, and P. Oittinen, "Stereoscopic depth perception in video see-through augmented reality within action space," *Journal of Electronic Imaging*, vol. 23, Jan-Feb 2014.
- [6] S. Reichelt, R. Haussler, G. Futterer, and N. Leister, "Depth cues in human visual perception and their realization in 3D displays," *Three-Dimensional Imaging, Visualization, and Display 2010 and Display Technologies and Applications for Defense, Security, and Avionics Iv*, vol. 7690, 2010.
- [7] P. Howard, *Perceiving in depth, volume 1: basic mechanisms*: Oxford University Press, 2012.
- [8] M. Banks, "The vertical horopter is not adaptable and is not adaptive for viewing along the ground," *I-Perception*, vol. 2, 2011.
- [9] K. N. Ogle, "Precision and validity of stereoscopic depth perception from double images," *J Opt Soc Am*, vol. 43, pp. 907-13, Oct 1953.
- [10] K. M. Schreiber, J. M. Hillis, H. R. Filippini, C. M. Schor, and M. S. Banks, "The surface of the empirical horopter," *Journal of Vision*, vol. 8, 2008.
- [11] S. Lamantia, D. Purves, G. J. Augustine, D. Fitzpatrick, and W. C. Hall, *NeuroScience: Sinauer Associates Incorporated*, 2011.
- [12] M. J. Tové, *An Introduction to the Visual System*: Cambridge University Press, 1996.
- [13] L. Xing, "Towards Reliable Stereoscopic 3D Quality Evaluation: Subjective Assessment and Objective Metrics," 2013.
- [14] M. Kytö, "Depth Perception of Augmented and Natural Scenes through Stereoscopic Systems," 2014.
- [15] M. Kyto, M. Nuutinen, and P. Oittinen, "Method for measuring stereo camera depth accuracy based on stereoscopic vision," *Three-Dimensional Imaging, Interaction, and Measurement*, vol. 7864, 2011.
- [16] State, J. Ackerman, G. Hirota, J. Lee, and H. Fuchs, "Dynamic virtual convergence for video see-through head-mounted displays: Maintaining maximum stereo overlap throughout a close-range work space," *Ieee and Acm International Symposium on Augmented Reality, Proceedings*, pp. 137-146, 2001.
- [17] T. Sielhorst, M. Feuerstein, and N. Navab, "Advanced Medical Displays: A Literature Review of Augmented Reality," *Journal of Display Technology*, vol. 4, pp. 451-467, Dec 2008.
- [18] State, M. A. Livingston, W. F. Garrett, G. Hirota, M. C. Whitton, E. D. Pisano, et al., "Technologies for augmented reality systems: realizing ultrasound-guided needle biopsies," presented at the Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996.
- [19] Bichlmeier, E. Euler, T. Blum, and N. Navab, "Evaluation of the virtual mirror as a navigational aid for augmented reality driven minimally invasive procedures," in *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, 2010, pp. 91-97.
- [20] H. Fuchs, M. A. Livingston, R. Raskar, D. Colucci, K. Keller, A. State, et al., "Augmented reality visualization for laparoscopic surgery," *Medical Image Computing and Computer-Assisted Intervention - Miccai'98*, vol. 1496, pp. 934-943, 1998.
- [21] V. Ferrari, G. Megali, E. Troia, A. Pietrabissa, and F. Mosca, "A 3-D mixed-reality system for stereoscopic visualization of medical dataset," *IEEE Trans Biomed Eng*, vol. 56, pp. 2627-33, Nov 2009.
- [22] Pietrabissa, L. Morelli, M. Ferrari, A. Peri, V. Ferrari, A. Moglia, et al., "Mixed reality for robotic treatment of a splenic artery aneurysm," *Surg Endosc*, vol. 24, p. 1204, May 2010.
- [23] Y. Abe, S. Sato, K. Kato, T. Hyakumachi, Y. Yanagibashi, M. Ito, et al., "A novel 3D guidance system using augmented reality for percutaneous vertebroplasty," *Journal of Neurosurgery-Spine*, vol. 19, pp. 492-501, Oct 2013.
- [24] F. Cutolo, S. Carli, P. D. Parchi, L. Canalini, M. Ferrari, M. Lisanti, et al., "AR interaction paradigm for closed reduction of long-bone fractures via external fixation," in *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, 2016, pp. 305-306.
- [25] F. Sauer, A. Khamene, B. Basclé, S. Vogt, and G. J. Rubino, "Augmented reality visualization in iMRI operating room: System description and pre-clinical testing," *Medical Imaging 2002: Visualization, Image-Guided Procedures, and Display*, vol. 4681, pp. 446-454, 2002.
- [26] G. Badiali, V. Ferrari, F. Cutolo, C. Freschi, D. Caramella, A. Bianchi, et al., "Augmented reality as an aid in maxillofacial surgery: Validation of a wearable system allowing maxillary repositioning," *Journal of Cranio-Maxillofacial Surgery*, vol. 42, pp. 1970-1976, Dec 2014.
- [27] F. Cutolo, G. Badiali, and V. Ferrari, "Human-PnP: Ergonomic AR Interaction Paradigm for Manual Placement of Rigid Bodies," in *Augmented Environments for Computer-Assisted Interventions*, ed: Springer International Publishing, 2015, pp. 50-60.
- [28] Ware, *Information visualization: perception for design*: Elsevier, 2012.
- [29] P. Milgram and M. Kruger, "Adaptation Effects in Stereo Due to Online Changes in Camera Configuration," *Stereoscopic Displays and Applications Iii*, vol. 1669, pp. 122-134, 1992.
- [30] Takagi, S. Yamazaki, Y. Saito, and N. Taniguchi, "Development of a stereo video see-through HMD for AR systems," *Ieee and Acm International Symposium on Augmented Reality, Proceeding*, pp. 68-77, 2000.
- [31] V. Ferrari, F. Cutolo, E. M. Calabro, and M. Ferrari, "HMD Video See Through AR with Unfixed Cameras Vergence," 2014 *Ieee International Symposium on Mixed and Augmented Reality (Ismar) - Science and Technology*, pp. 265-266, 2014.
- [32] F. Cutolo, P. D. Parchi, and V. Ferrari, "Video See Through AR Head-Mounted Display for Medical Procedures," 2014 *Ieee International Symposium on Mixed and Augmented Reality (Ismar) - Science and Technology*, pp. 393-396, 2014.
- [33] F. Cutolo, C. Freschi, S. Mascioli, P. Parchi, M. Ferrari, and V. Ferrari, "Robust and Accurate Algorithm for Wearable Stereoscopic Augmented Reality with Three Indistinguishable Markers," *Electronics*, vol. 5, p. 59, 2016.
- [34] F. Cutolo, M. Siesto, S. Mascioli, C. Freschi, M. Ferrari, and V. Ferrari, "Configurable Software Framework for 2D/3D Video See-Through Displays in Medical Applications," in *Augmented Reality, Virtual Reality, and Computer Graphics: Third International Conference, AVR 2016, Lecce, Italy, June 15-18, 2016. Proceedings, Part II*, L. T. De Paolis and A. Mongelli, Eds., ed Cham: Springer International Publishing, 2016, pp. 30-42.
- [35] S. Parrini, F. Cutolo, C. Freschi, M. Ferrari, and V. Ferrari, "Augmented reality system for freehand guide of magnetic endovascular devices," in 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 490-493.
- [36] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-Squares Fitting of 2 3-D Point Sets," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 699-700, Sep 1987.
- [37] K. Matsunaga, T. Yamamoto, K. Shidoji, and Y. Matsuki, "The effect of the ratio difference of overlapped areas of stereoscopic images on each eye in a teleoperation," *Stereoscopic Displays and Virtual Reality Systems Vii*, vol. 3957, pp. 236-243, 2000.

Comparison of Event Choreography and Orchestration Techniques in Microservice Architecture

Chaitanya K. Rudrabhatla
Executive Director - Solutions Architect
Media and Entertainment domain
Los Angeles, USA

Abstract—Microservice Architecture (MSA) is an architectural design pattern which was introduced to solve the challenges involved in achieving the horizontal scalability, high availability, modularity and infrastructure agility for the traditional monolithic applications. Though MSA comes with a large set of benefits, it is challenging to design isolated services using independent Database per Service pattern. We observed that with each micro service having its own database, when transactions span across multiple services, it becomes challenging to ensure data consistency across databases, particularly in case of roll backs. In case of monolithic applications using RDBMS databases, these distributed transactions and roll backs can be handled efficiently using 2 phase commit techniques. These techniques cannot be applied for isolated No-SQL databases in micro services. This research paper aims to address three things: 1) elucidate the challenges with distributed transactions and rollbacks in isolated No-SQL databases with dependent collections in MSA, 2) examine the application of event choreography and orchestration techniques for the Saga pattern implementation, and 3) present the fact-based recommendations on the saga pattern implementations for the use cases.

Keywords—Microservice architecture; database per service pattern; Saga pattern; orchestration; event choreography; No-SQL database; 2 phase commit

I. INTRODUCTION

According to Martin Flower, the microservice architectural style [2], [3] is an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API. MSA defines each service to be totally independent [4] with its own database. When MSA is defined with completely isolated No-SQL databases [6], and when the business transactions span across multiple services, the state changes in one database entity are not visible to state changes in the other. The application cannot use the local ACID transactions as the entities are now spread into multiple databases. Also, if the transaction gets rolled back because of a failure in one of the micro services, state recovery cannot be attained using the standard 2PC [8] as these are distributed entities. The scenario becomes even more challenging when there are dependent entities with one to many relationships.

To handle this scenario, saga pattern can be used [5]. The services which alter the state can be written in the form of a Saga. In a saga, each service which changes the state of the database in a distributed transaction [1], [11], can generate an event which can trigger the next micro service. In case of a failure, the saga triggers a sequence of compensating roll back events from one service to the other in the reverse direction. These sagas can be designed using two techniques: (1) Event choreography, in which each service can trigger other service's event without a central coordinator. (2) Orchestration, in which a central coordinator makes the decision of triggering the relevant events in the saga. Both these techniques have pros and cons based on the use case which is being implemented. In the past some researchers have suggested the use cases for which these approaches are suitable, but a quantitative analysis has not been performed. In this research, we tried to come up with the recommendations on which saga technique to pick up in which scenario by examining the performance and complexity using the factual data generated by simulating a variety of use cases using a custom project developed on spring boot based micro services and Mongo DB and ActiveMQ based java messaging service queue, which is explained in the later sections.

The rest of the paper is organized as follows. In Section II, we explain the challenges involved in the distributed transactions in the MSA with no-SQL databases by bringing up the use cases in an e-commerce application. In Section III, we explain how the event choreography and orchestration can be implemented for these use cases. In Section IV, we go through the relevant work conducted in the research project and outline the results. In Section V, the conclusions are presented.

II. DISTRIBUTED TRANSACTIONS IN MSA

In a traditional monolith application based on relational databases, the transactions originate and progress within the scope of the container hosting the application. So, it becomes easy to handle the roll backs. But it is different in case of micro services running with database per service pattern [7]. Since the entities and the databases are isolated, the traditional rollback approaches cannot be applied. We have taken an example of a standard e-commerce application flow (Fig. 1) to explain the complexity of distributed transactions.

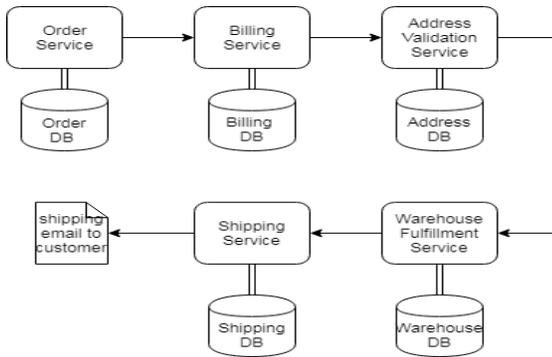


Fig. 1. Micro services in an e-commerce application.

As it can be seen, the order placement, credit card billing, address validation, fulfillment and inventory update, shipping are the various micro services which have their own databases and entities. It is not possible to capture all the steps in a single ACID transaction. To ensure the data consistency [13], we need to implement distributed transaction. Since there is no direct linking of the entities or databases, when the distributed transaction progresses few steps and encounters an issue, it becomes challenging to handle the consistency in the entity states by performing the roll backs. For example, when an order is placed successfully, and the customer’s credit card is charged, but if the address validation fails, the transaction must be rolled back correctly so that the customer is not charged for the unfulfilled item. That means the transaction must be rolled back in the proper reverse order. To handle this flow of events in forward and reserve directions by triggering the relevant events, Saga pattern can be used. Saga pattern can be implemented using Event choreography and orchestration techniques as mentioned below.

III. EVENT CHOREOGRAPHY VS. ORCHESTRATION

Some researchers already explored how event choreography and Orchestration [12] techniques for implementing sagas in micro service architecture. We are going to explain it in detail with the use case of e-commerce application mentioned above. In Event choreography approach, when a micro service executes a local transaction, it publishes an event which can be subscribed by one or other micro services to trigger their local transactions. This process proceeds till the last service which doesn’t publish any more events, there by marking the end of transaction. It can be visualized in Fig. 2 given below.

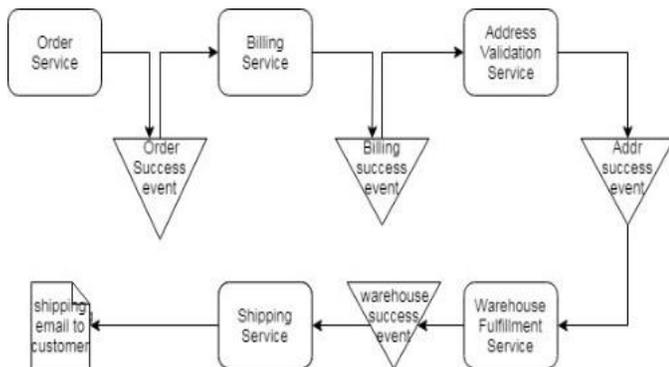


Fig. 2. Event choreography flow.

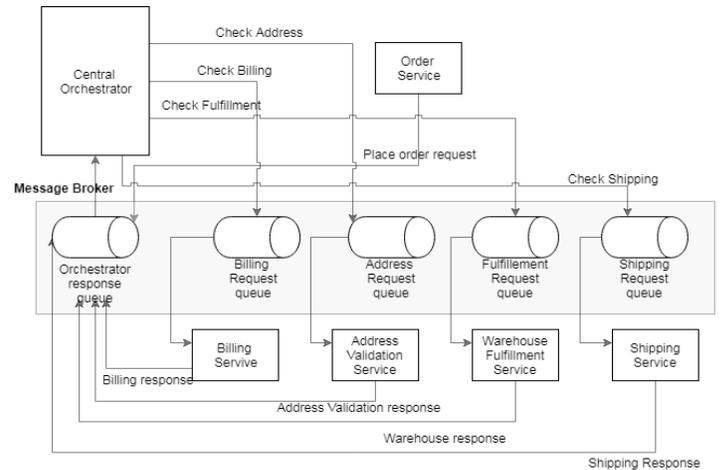


Fig. 3. Orchestration flow.

In this approach there is no central coordinator which listens to the events and triggers the relevant micro service local transaction.

The other technique to implement sagas is called orchestration. In this, there is a central coordinator which listens to all the events emitted by any of the micro service local transaction. Based on the incoming event, it triggers the next local transaction in a different micro service or services. This pattern is depicted in Fig. 3 below.

The scenarios mentioned above are depicted using a single entity at each local transaction level. It can be complex if there are dependent collections in each of those data sources. When a transaction needs to be rolled back, the dependent collections state needs to be reverted as well. Both the techniques mentioned above have pros and cons based on the scenario that needs to be implemented. In the next section, we are going to simulate various scenarios and understand the suitability of these techniques.

IV. RELATED WORK: RESEARCH PROJECT

To determine which saga implementation technique is more suitable under which scenario, we have implemented a research project and simulated various circumstances. We have implemented micro services in spring boot technology. A service discovery component called Eureka [9], [14] is used to register and discover the micro services running. This is similar to the other API gateways like Kong or Apigee which are available in the market. The entities are represented as collections in an open source no-SQL database called Mongo. Each micro service -MS1, MS2, MSn has an isolated instance of Mongo DB- DB1, DB2, DBn, respectively with a collection -C1, C2, Cn, respectively running on each of those database instances. These micro services and database instances run on Linux based virtual machines. First the event choreography technique is executed with 2 micro services, MS1 and MS2 having DB1 and DB2 as databases for each micro service with C1 and C2 as collections in each database respectively as depicted in Fig. 4. Each collection has an attribute called state which describes the state of the entity with the possible values of S1 and S2 and an attribute called timestamp which records the time stamp when the state change occurred.

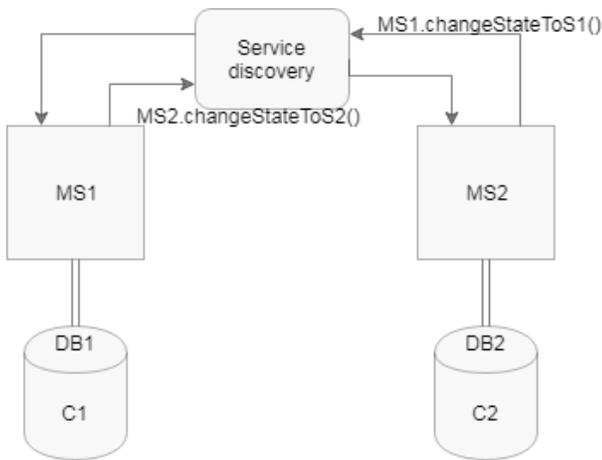


Fig. 4. Event choreography with 2 micro services.

A. Performance Analysis

Here is the sequence of steps which are executed as a part of the project to compare the performances:

- Micro service MS1 method is called which changes the value of state attribute of collection C1 from 'S1' to 'S2'. This method saves the time stamp T1 of the update action in the timestamp attribute of C1.
- Once the update is complete, MS1 triggers an event called 'MS1_state_change_success' which calls the method 'changeStateToS2' on micro service MS2.
- MS2 executes a logic to update state of C2 from 'S1' to 'S2'. But we simulate the transaction failure with which the state change of C2 fails.
- Now due to transaction failure, MS2 creates an event called 'MS2_state_change_failure' which rolls back the transaction in MS2 and calls the method 'changeStateToS1' on micro service MS1.
- MS1 then rolls back the state of C1 from 'S2' and 'S1' and updates the time stamp to new value T2.
- The difference between T2 and T1 tells us the time taken to execute the Saga with Event choreography of 2 micro services. These values are noted down as time taken for 2 micro service event choreography.
- Similarly, this exercise is repeated 3 more times by taking 4 micro services, 6 micro services and 8 micro services in each attempt. The exercise is executed in the same fashion as described in the steps above where the transaction progresses in a series of events from MS1 to MSn-1. At MSn-1 it triggers the event 'MSn-1_state_change_success' and calls the 'changeStateToS2' method on MSn. MSn fails the transaction and rolls back the transaction by calling the 'changeStateToS1' on MSn-1. This rolls back the state of Cn-1 to S1 and triggers the method 'changeStateToS1' on MSn-2. This happens till it reaches 'changeStateToS1' on MS1 which rolls back the state to S1 and calculates the time difference.

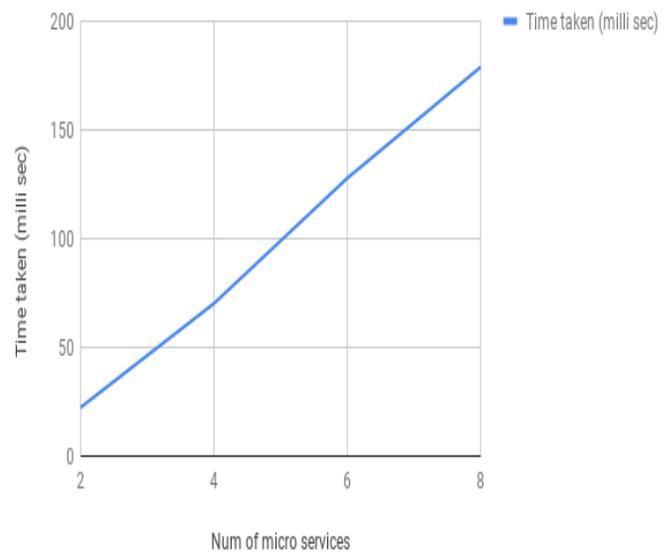


Fig. 5. Correlation of time taken vs. Micro services in event choreography.

We have executed 5 test runs and noted down the time taken in each instance and calculated the average. The graph in Fig. 5 shows the time taken vs number of microservices in the event choreography.

A similar exercise is performed using orchestration technique. In this, a central orchestration service is added which listens to various events and takes the necessary action. We have used an Apache ActiveMQ [10] as the JMS broker. Here is the sequence of steps which take place.

- MS1, Ms2 MSn are the microservices, each having a mongo DB instance DB1, DB2 DBn. Each of the databases has collections C1, C2 Cn. Like the setup described in event choreography.
- For Orchestration technique we hosted a new micro service MSn+1.
- We have n different queues running on Active MQ broker Q2, Q3...Qn+1 with MS2, MS3...MSn+1 subscribing to each of them, respectively.
- When state change happens from S1 to S2 on MS1, it triggers an event 'MS1_state_change_success' on the orchestrator MSn+1.
- Orchestrator posts a message on Q2, which MS2 listens and executes 'changeStateToS2' method and changes state to S2. Upon state change, MS2 posts a message 'MS2_state_change_success' on the Qn+1 which is subscribed orchestrator MSn+1.
- This forward transaction continues till it reaches the last micro service MSn. At MSn we fail the transaction, roll back the state to S1 on Cn and post the message 'MSn_state_change_failure' on the Qn+1 which is subscribed orchestrator MSn+1.
- Orchestrator listens to this roll back event from MSn and posts a rollback message on Qn-1. MSn-1 listens to this message and rolls back the state to S1 on MSn-1.

- This rollback continues till it reaches MS1 which rolls backs the state to S1, notes the time difference and posts no more messages.
- This exercise is also performed 4 times, with 2,4,6,8 micro services and orchestrator and the timestamps are noted.

In Fig. 6 given below, the graph shows the time taken vs number of microservices in the orchestration technique.

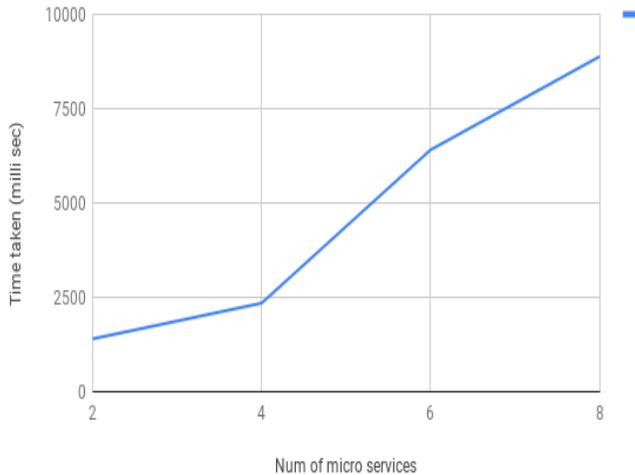


Fig. 6. Correlation of time taken vs. Micro services in orchestration.

Now it can be clearly seen that event choreography takes much faster in performance when compared to Orchestration. Event choreography can be well suited in the scenarios where the number of micro service calls are limited, and the response time is critical.

B. Complexity

The same experiment is repeated with a scenario which is more complex. When a state change is occurred in one micro service, we want to test both the techniques by triggering multiple events in more than one micro service. To do this, we implemented the following pattern.

- When state changes from S1 to S2 in C1, MS1 triggers 2 events which changes the state of C2 in MS2 from S1 to S2 and state of C3 in MS3 from S1 to S3.
- Upon successful change of state from S1 to S2 in C2, MS2 triggers 2 events which changes the state of C3 in MS3 from S3 to S2 and state of C4 in MS4 from S1 to S3.
- When a transaction fails at MS4, it rolls back the state of C4 to S1 and triggers 2 events which changes the state of C3 in MS3 from S2 to S1 and state of C2 in MS2 from S2 to S3.
- Upon successful rollback of state from S2 to S1 in C3, MS3 triggers 2 events which changes the state of C2 in MS2 from S3 to S1 and state of C1 in MS1 from S2 to S3.

- Finally, upon successful rollback of state from S3 to S1 in C2, MS2 triggers an event which changes the state of C1 in MS1 from S3 to S1.

This pattern is performed for 4 micro services and 6 micro services in both event choreography and orchestration for 5 test runs. It was observed that the time taken for orchestration technique is approximately 40 times more than the event choreography. But it was noted that as the number of events increased, it became more and more complex to handle the code in individual micro services. Whereas orchestrator proved to be more elegant in handling multiple events with less confusion as the event handlers are orchestrator at a single location.

C. Load based Test

The same setup is repeated one more time with a scenario where the frequency of events which are triggered are increased by 5-fold and 10-fold. This is obtained by writing a test client which fires parallel requests. We calculated the ratio of response times with the frequency of 1 vs 5 vs 10. We observed that the event model began to respond slowly as the frequency increased, whereas the orchestrator was able to handle the load better. The response times varied as 1:3.6:8.2 for event, whereas the ratios for orchestration came out as 1:3.9:6.4. These results might have been different if we ran multiple instances of each micro service rather than a single instance by horizontally scaling them using auto scaling techniques available in the cloud. This can be an element of future research.

V. CONCLUSION

In this paper, we performed a quantitative analysis of performance of both event choreography and orchestration techniques used for implementing the saga design pattern to handle the distributed transactions in isolated no-SQL databases in micro service architecture. We were able to clearly determine that event choreography is much faster in performance when compared to orchestration. However, event choreography becomes very complex to code and handle if there are multiple events triggered from each micro service. It is also evident that handling multiple actions for the triggers without a central orchestrator is tough as one developer or team working on a micro service may not be aware of the other. This shows that event choreography is a suggested approach when there are less number of micro services participating in the distributed transaction, or the number of event triggers are not too many or when the trigger actions are not too complex. Orchestration is slow, but it is useful when the transaction scenarios are complex.

Future work includes working on scenarios involving transaction rollbacks involving dependent collections where the states are distributed in multiple collections and recording the performance metrics in various saga patterns. We also plan to do research around the areas where the triggered actions are bi-directional or cyclic rather than unidirectional and record the metrics around them. Author is thankful to anonymous reviewers for their valuable feedback.

REFERENCES

- [1] R.K. Batra, M. Rusinkiewicz & D. Georgakopoulos, A decentralised deadlock-free concurrency control method for multidatabase transactions, Proc. 12th Int. Conf. on Distributed Computing Systems, 1992.
- [2] N. Alshuqayran - A Systematic Mapping Study in Microservice Architecture. In Proc. of the 9th International Conference on Service-Oriented Computing and Applications. IEEE, IEEE, 2016.
- [3] Paolo Di Francesco- Architecting Microservices. 2017 IEEE International Conference on Software Architecture Workshops.
- [4] H. Kang, M . Le, and S. Tao, "Container and microservice driven design for cloud infrastructure DevOps," in 2 0 1 6 IEEE International Conference on Cloud Engineering (IC2E). IEEE, 2 0 1 6 , pp. 202-2 1 1.
- [5] Hector Garcia-Molina, Kenneth Salem - Proceedings of the 1987 ACM SIGMOD international conference on Management of data, pages 249-259. <https://dl.acm.org/citation.cfm?id=38742>
- [6] A.K. Elmagarmid & W. Du, A paradigm for concurrency control in heterogeneous distributed database systems, Proc. 6th Int. Conf. on Data Engineering, 1990.
- [7] Messina, Antonio & Rizzo, Riccardo & Storniolo, Pietro & Tripiciano, Mario & Urso, Alfonso. (2016). The Database-is-the-Service Pattern for Microservice Architectures. 9832. 223-233. 10.1007/978-3-319-43949-5_18.
- [8] 2 phase commit - https://en.wikipedia.org/wiki/Two-phase_commit_protocol.
- [9] P. Bak, R . Melamed, D . Moshkovich, Y . Nardi, H. Ship, and A . Yaeli, "Location and context-based microservices for mobile and internet of things workloads," in 2 0 1 5 IEEE International Conference on Mobile Services, 2 0 1 5 , pp. 1-8.
- [10] Apache Active MQ JMS - <http://activemq.apache.org/>
- [11] P.A. Bernstein & N. Goodman, Concurrency control in distributed database systems, AGM Computing Surveys 13(2) pp. 185-222, 1981.
- [12] J. P. Macker and I. Taylor. Orchestration and analysis of decentralized workflows within heterogeneous networking infrastructures. Future Generation Computer Systems, 2017.
- [13] J. Tang, Using dummy reads to maintain consistency in heterogeneous database systems, Proc. Third Workshop on Future Trends of Distributed Computing Systems, 1992.
- [14] Tasneem Salah, M. Jamal Zemerly, Chan Yeob Yeun, Mahmoud Al-Qutayri, Yousof Al-Hammadi. The Evolution of Distributed Systems Towards Microservices Architecture. The 11th International Conference for Internet Technology and Secured Transactions (ICITST-2016)

Location-based E-Commerce Services: (Re-) Designing using the ISO9126 Standard

Antonia Stefani, Bill Vassiliadis, Theofanis Efthimiades
Hellenic Open University
Patras, Greece

Abstract—E-commerce services based on user geographic location have emerged as a particularly important segment of modern information services. In these user-intensive applications, quality of service is important and design methods are increasingly relying on software standards to achieve quality. In this paper, we propose an evaluation model for location based e-services that provide insights on how overall system quality can be strengthened via identifying the most important quality characteristics of specific user-system interactions facets. The model categorizes location based services into taxonomies of components / functions, which are further analyzed in interaction facets and significance levels. A further mapping to external qualitative sub-characteristics of the ISO9126 quality standard is used to formally decompose design quality into quality attributes. The view of software design through quality attributes is supported by a mathematical model, which calculates significance weights on service components, defined either by designers or by the end users. An experiment, where this method is used to assess functionality is presented.

Keywords—E-commerce; location based services; software quality; software design; ISO9126

I. INTRODUCTION

Mobile commerce has grown rapidly in recent years as infrastructure, hardware, and software-supporting technology has dramatically improved its speed and reliability. Mobile devices now offer a plethora of services based on push/pull models of information based on user location. Their energy autonomy and processing power are no longer such serious obstacles as they were in the past and developers have the flexibility to develop resource-demanding software that has in turn, greatly contributed to mobile apps success. Especially mobile e-commerce software offers two unique benefits. It may use either apps or lightly-tailored browsers using well-known user-software interaction patterns, as well as location based services (LBS) that geographically link the real to the virtual world [1].

Competition from e-commerce and e-services vendors has led to the offerings of systems with an ever-increasing complexity. Functional and non-functional requirements recognised at the early design stages of software development are largely based on user expectations and define crucial architectural design decisions. Achieving quality of service, one of the competitive advantages of a modern e-services vendors, depends on the quality performance of specific architectural properties such as functionality, reliability, usability, just to mention a few [2]. In order to make the correct much-needed architectural decisions at the early stages of

system design, a certain level of confidence to the results of these decisions is appropriate. One way to achieve this goal is to rely on formal software standards [3].

The quality models defined in ISO standards such as ISO9126 and ISO25010 decompose software quality into characteristics organised into a hierarchical structure in order to facilitate the establishment of requirements and general criteria for their satisfaction [4], [5]. Design quality can therefore be addressed in the terms of how (and how much) quality characteristics influence software components. Targeted design can be achieved by taking decisions that favour the quality of certain components (that may be of most value to end-users) over others. Measurement of quality characteristics, where possible, is valuable to battle against the generality of formal standards and increase practical impact [6].

In this work we propose a Quality Model for designing LBS as sub-systems to e-commerce services. This is a research subject that poses several difficulties in relation to other on-line software, namely the push/pull model of information flow, the interaction with objects based on spatial proximity to the user, managing layers of dynamic information and different interaction facets between users and services [7]. Capturing user requirements and taking design or re-design decisions through evaluation, is one method that ensures user participation in the development process [8], [9].

In order to provide a clearer view of which software components need to be evaluated, the LBS sub-system is analysed in functions/components. The ISO 9126 software evaluation standard was used as the basis for the qualitative assessment, a general standard that can be applied to such systems but has not been widely used in e-commerce business-type systems to date [10]. A qualitative connection of these components to the characteristics of ISO 9126 is possible in order to pinpoint how each component should be assessed. The result of this step is the definition of a model - quality map of the LBS subsystem [11], [12].

Further refinement of the quality design process includes the categorization of functions into significance levels and facets of user-software interaction. Three levels of significance are identified. The first level includes functions that are considered essential to the user and are technology-independent (e.g. locating an address). The second level contains important functionality, which is considered desirable from the majority of users (e.g. focus on map points). Finally, the 3rd level includes functions that are more dependent on the technology used (e.g. road mapping functions). Then the

functions are categorized into five viewpoints (aspects of user interaction): Presentation, Navigation, Routing, Information and Purchasing. For each characteristic of the system, a correlation function is attributed to depict its relation to the qualitative sub-characteristics of the model. The model defines significance weights for each level, qualitative feature and sub-feature. The proposed methodology and the mathematical model that complements it assign weight to the characteristics of the functions/components of the LBS sub-system in order to organize system requirements according to end-user preferences. The later can be used for system re-design or for the design of new systems/sub-systems. The Quality Model can also be used for guided system development since expert/designers may set pre-defined values to model quality parameters and derive appropriate quality requirements. An experiment was conducted for calculating these weights and provides insights on how to use the proposed model.

This paper is structured as follows: in Section II, the basic principles of the Quality Model are presented while in Section III the mapping process of system attributes to the quality characteristics of ISO9126 is explained. The mathematical model used for calculating the correlation between system functions and quality model attributes is also presented. An experiment showcases the application of the method in Section IV and conclusions are drawn in Section V.

II. QUALITY MODEL

A. Rational and Structure

The software subsystem (LBS of e-commerce system) is analysed into a set of basic functions. These functions are either explicitly mentioned in the requirements document and/or may include functions that are desirable (to be included in the system). It is the case in many software development projects that desirable configurations are either not possible to achieve due to time or budget constraints or may not be actually popular with users [13]. A software's added value, as viewed in general quality management principles is increased, theoretically, with the number of (new) features it possesses. In the case of e-services, features usually correspond to functionality. There is a tendency to design services with many functions, however, setting a goal for quality over a large number of functions poses a stress to project resources and project management [2]. There is a need for a structured and organised method to achieve quality over interdependent functionality taking into account user expectations. Targeted quality design helps designers to better understand how different system functionality influences overall system quality or even allows them to adjust the design so as to achieve a certain degree of influence. This adjustment may come as a result of project development limitations, special requirements by specific target groups, application of agile methods to software development or technical limitations [14]. The quality of the different parts of the system should also be influenced by user expectations i.e. of what is considered qualitative for which component. The quality map of the system should also address the problem of how the overall quality of the system is influenced by its components. To answer this question, one needs to identify the components and evaluate their contribution to quality. This is mainly answered by the users;

they are usually mainly concerned with the set of available functions (addressed by the Functionality quality characteristic of ISO9126) and their quality (addressed by Quality in Use) [3], [15]. We extend the model presented in [11] to include a more comprehensive link between the processes of software analysis and design and the production and use of the system quality map. We further configure the model to address LBS specific quality issues.

Using a divide and conquer strategy, the services are analysed in their basic functionality (during the analysis phase of the software development lifecycle). They are further organised into significance levels and facets of user-system interactions. Significance levels are useful for incorporating a ranking of services importance, a prioritization mainly derived from the designer team knowledge of the business and technical context in which the services will operate. Prioritization also helps achieve economy of scale where needed resources are not timely available or not available at all. Facets further organise functions into categories of system-user interaction taxonomies where the type of interaction (and not significance per se) is considered. User perception of quality is introduced by calculating weights that quantify the contribution of significance of each function to the overall system quality. This is a user's view of the system quality. It is further detailed by the mapping of functions to quality characteristics and sub-characteristics of ISO9126 and the assignment of weights to the mapping relations. The model permits the specific targeting of quality sub-characteristics for each function (setting quality sub-goals). Strong relationships, that is high values of a weight for function A to quality sub-characteristics x, means that the development team should take specific steps to reach this goal. The nature of the sub-characteristic itself provides general guidelines on what is considered qualitative. Quality sub-goals are set depending on the resources available, the technology used, the experience and knowledge and of the quality culture of the development team. To this end, either the top two weights (as per value) for each function may be considered or a cut-off value to indicate whether a mapping relationship is strong or weak. Strong relationships help define general design goals (global quality goals), a process somewhat not straightforward. Trade-offs surely exists between sub-goals depending on technological and/or methodological factors.

The process that derives the system quality map (the Quality Model) is depicted in Fig. 1. The phase of analysis (system breakdown to functions), taxonomy build-up (organisation of functions according to significance level and facets) corresponds to analysis tasks of the software development lifecycle. Calculation of weights and mapping correspond to early design tasks. The weighting phase requires the gathering of knowledge of how users perceive software quality. The gathering process must take place either during the analysis of requirements using methods such as user surveys, benchmarking, expert reviews and by taking advantage of the corporate knowledge in the specific context of use. Specifically for LBS, three significance levels and four main facets are considered in this process (a fifth facet is considered not mandatory).

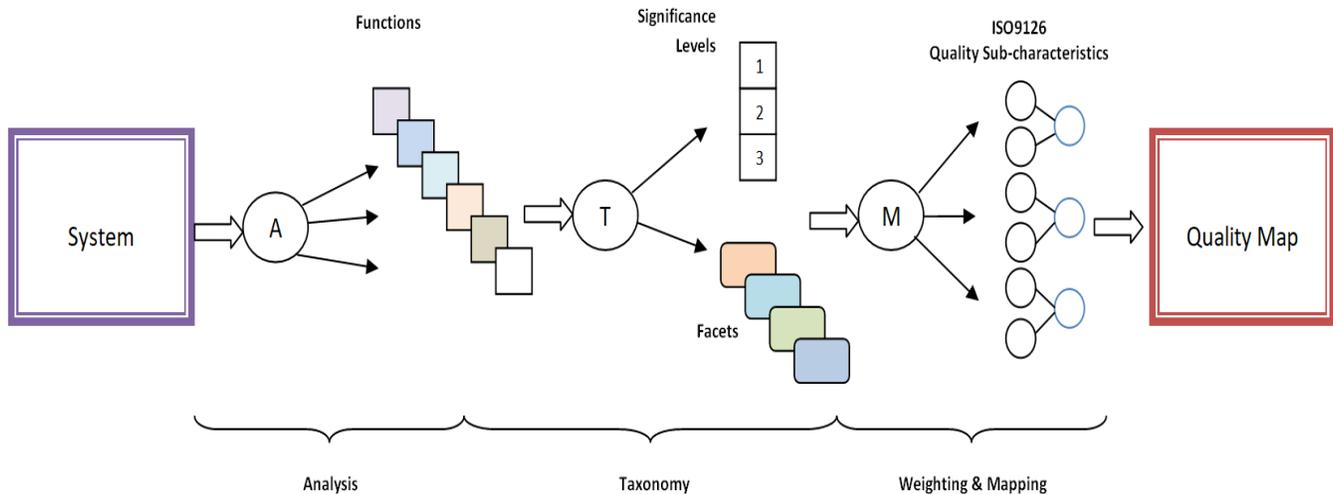


Fig. 1. The process that produces the quality map.

TABLE I. LOCATION BASED SERVICES FUNCTIONS

Service	Description
Map view/ My Location	Location of the user in a map is presented with an accuracy of a few meters based on GPS or Wi-Fi data.
Point of Interest (POI)	Search and locate any Points of Interest (POI) close to the user location or in a broader geographical area.
Directions (routing)	Providing directions from a starting point to a destination using various means of transportation. It is possible to use the current position of the user or any other persons as the starting point, by entering the postal address or by selecting a point on the map.
Locate Friends	Find friends located nearby and communicate with them using social networking applications.
Mode (Transit and/or walking directions)	Information on routes for the user to travel by bus or train, as well as provision of walking directions. This feature requires information from transport organizations, who have to update the system in real time.
Street view	A 3D visual feature with pictures of the actual road, as it was captured at an earlier time. Through the mobile phone, the user can see the real image of a selected point of interest, as well as have a more general view of the street in which it is located.
Traffic	A real-time update of traffic conditions on user-selected roads, providing assistance in choosing the fastest route.
Apps connection	The ability to wirelessly forward emails, calendar entries, and phonebooks from the sub-system to the e-commerce system or an external app

The use of appropriate quality mechanisms to assess the external quality of the system (quality as it is perceived by the final users) is an important objective in each evaluation. Evaluation methods deliver better results when used to evaluate specific components of the system by prioritizing a qualitative goal [2]. The interpretation of measurements and the contribution of quality sub-characteristics to the assessment of the quality of the system are the elements that determine its mapping to software functions. The need to select the appropriate mapping (function to quality sub-characteristic) for the assessment of each component can be viewed in terms of a user-centred approach that satisfies basic principles of quality in use: how the software behaves when operated by the users. In order to evaluate the external quality, interaction facets (similar to the stakeholder viewpoints of a system architecture in ISO42010: 2011 [16]) were applied so as to better identify how the system responds to user actions. A set of basic LBS functions are depicted in Table I. These services are further analysed into functions, which correspond to specific facets.

B. The Navigation Facet

The Navigation facet includes mechanisms that support user navigation on a map. The interaction between the user and the LBS is either in a one-dimensional mode (for example, using a stylus on a touch screen) or in a multidimensional mode (for example, a combination of voice and motion). The user's navigation capabilities in the application generally include the following features:

- a) *Move within map*: The user may gradually move within the map in different directions by moving the stylus on the touch screen or navigating to the appropriate menu, displaying the areas of interest.
- b) *Focusing on map points or areas*: The user can focus on map points or expand areas without changing the content of the displayed information. The areas shown in the screen are either stored off-line or they are downloaded in real-time.
- c) *Hide / Show Points*: The user can hide or highlight certain objects or classes of objects (friends or points of

interest) displayed on the map by selecting them with the stylus or the appropriate combination of keys.

d) Analysis of points of interest: The user chooses a point of interest and retrieves information about it from the corresponding database maintained by the data provider.

e) Viewing settings: The user can change the viewing angle and parameters related to the graphical display such as colour, font of texts and symbols, terrain map contrast (day / night mode), display dimensions (2D or 3D), defining the values of the variables within the allowed limits.

f) Dialog window: The user, through the activation of dialogs, introduces his / her preferences and interests by changing the way it interacts with the application (e.g. chooses to disable voice instructions).

g) Search: Based on one or more criteria, the user can search for comparable entities or POIs by using the appropriate dialog box.

h) Shortcuts: Many application functions can be executed via the alphanumeric keypad, since the selection of certain keys corresponds to a specific action performed by the app.

i) Help: The user can learn about the way the application works.

C. The Routing Facet

The Routing facet includes mechanisms and methods that serve the user and system interaction when using map navigation services. The basic requirement of a user of a platform that implements map navigation services is to find a way of moving from one point to another, by deciding on the best route from a set of alternatives. The main parts of a route are the starting point, the various legs of the route and the destination. The starting point and the destination are determined by the following mechanisms:

a) My location: The current location of the user, as calculated by the positioning system (Wifi, GSM, GPRS or GPS), may be the starting point or destination of the route.

b) Point selection: The user selects the point on the map from which a route will start or end.

c) Point Search: Using the appropriate dialog, the user enters the desired mailing address to be used as the starting point or destination.

d) History: Points searched by the user in chronological order, starting from the most recent, can be the starting point or destination of the route.

e) Favourites: By storing "favourite" points on the map, the user can use them by referring to the appropriate list.

f) Friend position: The location of a friend identified by the system is a possible starting point or final destination of a route.

Concerning the route per se, the following methods of interaction are usually provided:

g) Direction: By taking into account the selected start and arrival points selected by the user, the application, computes the route, showing it on the map, highlighting the

nodal points and providing information identifying it. It also creates a list of detailed directions from the starting point, the intermediate points, to the final destination.

h) Transit: The user is informed about the time required to reach the destination using a particular route and based on the start time, calculates the time of arrival at the destination depending on the transport means selected. It also depicts the exact time each hub will be reached, based on traffic conditions, and if necessary, modify the route to be followed.

i) Walking: The route between two points can be determined, taking into account that the user has selected the pedestrian mode. In this case, the routing is not affected by road traffic and the shorter path is calculated more directly.

j) Reverse: The starting point and the destination can be reversed.

D. The Presentation Facet

Presentation includes the functions the system uses to present information to the user, including area maps. The goal of this facet is the comprehensive presentation of information using images, three-dimensional representations and sound. In an application that implements map navigation services, the user interface must provide the user with specific information, such as static map images of the selected area, user location on the map information about points of interest (friends or mobile objects), as well as directions on a route. A navigation service platform is required to present the requested information in the following ways:

a) Maps: Graphical representation of a geographical area using a road or terrestrial / satellite map. The map may include various information elements: (i) the route drawn from the selected starting point or the current location of the user to the selected destination; (ii) the points of interest and the road using three-dimensional graphics; (iii) colouring strongly specific streets depending on information such as traffic and (iv) the location of the user and that of his friends as well as the location of various entities / objects (such as the home or the car of the user).

b) Photos: View photos of roads and points of interest.

c) Text: Use of text to describe a point of interest (postal address, historical data), user status (speed, altitude, geographic coordinates of its location, if it is pedestrian or not) and a selected route, indicating route directions, intermediate and endpoints.

d) Voice: Use of voice instructions to allow the user to navigate a route or enter the route itself (starting point and destination).

E. The Information Facet

This facet includes the methods by which the user is aware of information other than the design and implementation of a route and the mechanisms by which this information affects the user interface. Besides the user's basic requirement for route creation, an essential feature is also the information about entities and situations that are located upon the area where the route is situated. The key elements of the information provided are the following:

a) *Update placement of moving / fixed entities*: The user navigating through the geographical map of the area or through the list of POIs is informed about the location of friends and objects of interest.

b) *Traffic update*: The user is informed about traffic conditions and special events (e.g. traffic jams) in selected roads by referring to the corresponding map.

c) *Route / Billing Update*: The user is informed about the potential charges (e.g. tolls) on various routes.

d) *POI related information*: By selecting a point of interest from a search result, the user is informed about features such as its location on the map, its postal address and various contact details (phone, e-mail, web site), the distance from public transport hubs and routes, the type and cost of the services it provides, as well as user comments. In addition, it is possible to view photos of the point of interest and the streets where it is located.

F. The Purchasing Facet

This optional facet supports B2C (business to Consumer) and G2C (Government to Consumer) processes. Points of Sales (PoS) are located through various push/pull modes. In pull mode, PoS are located in the map (in the area or point of interest) or they appear in the results of a user search. In push mode, a PoS appears in the map when the user enters an area geographically close to it. In push mode, the LBS sub-system can be configured to include/exclude objects based on user preferences, friends' recommendations or based on the existence of special offers.

The user goes through the following steps to make a purchase: i) location of PoS (push/pull mode), ii) purchase of service or product either via the parent application or by accessing an external, app or web store and iii. delivery of the purchased service. The facet includes features that support directly the purchasing process. Features that are related to navigation are not included (e.g. browsing through a list of PoS or services).

G. Levels of Significance

The quality model uses three levels of significance: the first level of significance includes, by definition, the functions that support the user regardless of the interaction medium. The 2nd level contains functions that the user desires. Finally, the third level includes the functions associated with the technological characteristics of the system. The three-tier structure allows the Quality Model to be expanded so as to be applicable to related or new, evolving systems, such as virtual reality imaging systems.

Level 1 includes those components and functions, which must be included in each navigation application, and their existence is independent of the technology used to support the system. Functions of this level include: 1) my location, 2) the point search on the map, and 3) the route along with routing

directions. The basic requirement of a user navigating a map is to acquire information on the various routes a selected point can be reached and how. For this reason, the integration and implementation of the above functions is crucial for the system's success and for quality assurance.

The 2nd level includes functions that are not that basic but are desirable by the user (or necessary to be included in the sub-system by the designers). They may be incorporated into the sub-system without requiring special technological specifications. Their incorporation to the system however does increase overall system quality. Additional functionalities and services at this level enhance user and system interaction. The basic user requirement satisfied in this level includes functions that inform the user of what objects are located around him, and additional information about those objects. Appropriate mechanisms to facilitate these interactions are used such as 1) history, 2) favourites, 3) walking, 4) inversion, 5) location of fixed entities and 6) information about points of interest.

Level 3 includes advanced operations, which depend on the hardware and software technology that the application uses to implement it. Technology dependence refers to not only software and hardware compatibility but to performance factors as well. At this level, functions correspond to user requirements for advanced product features and their implementation requires the use of advanced networks and devices capable of utilizing fast data processing and storage systems. The functions included at this level are: 1) 3D representation of routes or POS, 2) finding a friend (requires installation of specific software by all parties involved), 3) real-time updates on transit using public transport (requires connection between the provider's network and the transport operator), 4) real-time updates about traffic information (requires connection to traffic management centres G2C services) 5) information on routes and charges for public transport (requires a connection between the provider's network and the transport operator). Usually, traffic and billing information are not provided in real time but correspond to known patterns (e.g. heavy traffic is expected in main city routes in early morning hours) or somewhat out of date information from other sources.

III. MAPPING PROCESS

A. Mapping Functions to ISO9126 Characteristics

The Quality Model maps the functions and components of the system to the external characteristics and sub- of the ISO 9126 quality standard. The standard includes four quality characteristics (Table II).

The above-mentioned attributes determine the end user's view of the features and services provided by the system and can be used when interacting with it. ISO25010 [5] provides a similar, albeit more detailed, classification of characteristics/sub-characteristics that still remain to be tested for their practical value [17].

TABLE II. ISO9126 KEY QUALITY CHARACTERISTICS

Quality Characteristic	Description
Functionality	The ability of the software to provide all the required functions under defined and real conditions. It refers to the definition of the functions that the software should provide to meet user needs. The significance of the above feature is summarized in the question: "What are the functions of the software that meet user needs?".
Reliability	The ability of the software to operate in a fixed and specific manner under defined conditions.
Efficiency	The ability of the software to operate under defined conditions.
Usability	The characteristic of the software of being easy to use. Ease of use can be measured by how quickly a task is performed, how many mistakes are made, how quickly the software is learned and how satisfied final users are when using it.

B. The Mapping Function

A key element of the Quality Model is the definition of the correlation between the components/functions of the system with the external qualitative sub-features of the ISO9126:

$$(Component) \times (Quality Feature) \tag{1}$$

This formula creates a two-dimensional table for each one of the three levels of significance defined previously. The correlation between a component/function of the system with a particular sub-feature is determined by the correlation function (CF) r_{ij} (where i defines the component for each interaction facet and j defines the qualitative sub-characteristic of ISO9126). The set of values for the function is binary, either zero or one. If there is no (strong) correlation between a component and a sub-characteristic, then the function value is zero (denoted by 'O' in the table), otherwise it is '1' (denoted by 'X' in the table). The value of CF, and consequently the correlation between the two parameters in the table that is formed, is determined by expert evaluators or system designers. It can be redefined when a new function/component is introduced or it can be set so as to depict a quality design goal.

C. Mapping Table: The Functionality Quality Characteristic

Tables III, IV and V depict the functions/components of the three facets grouped in the significance levels 1, 2 and 3, as well as the correlation with the qualitative sub-characteristic of Functionality.

The Quality feature of Functionality refers to the set of functions that support user-system communication. Similar mapping takes places for the remaining three characteristics of ISO9126 (Reliability, Usability and Efficiency). Actually, the mapping process is compatible with all ISO hierarchical standards. The functions provided need to meet the needs and requirements of the user by implementing navigation, retrieval and processing of data and information. The realisation of these functions answers the question of 'what is the user able to do when interacting with the system?' The user, generally has the ability to use the service through text, images, audio, 3D imaging and graphics. The two key elements through which the user accesses the information are maps and text. Sophisticated devices also provide visualization and voice guidance mechanisms, enhancing application functionality and increasing the range of services offered. In each software application, including map navigation applications, the way the user navigates through menus and pages is a key element in assessing the quality of each software system whether it uses the web to implement it or not.

TABLE III. 1ST LEVEL MAPPING FOR FUNCTIONALITY

Functional Suitability				
Level 1	Sub-characteristic			
	Suitability	Precision	Interoperability	Security
Correlation r_{ij}				
Presentation Facet				
Maps (routing)	X	X	X	O
Maps (user position)	X	X	X	X
Text info (info tags)	X	X	X	O
Navigation Facet				
Move within map	X	X	O	O
Focus	X	X	O	O
2D presentation	X	O	X	O
Dialog Screens	X	O	X	X
Help	X	O	O	O
Routing Facet				
My position	X	X	X	X
Search for POI/PoS	X	X	X	O
Show route(s)	X	X	X	O
Information Facet				
POI information	X	X	X	O
Purchasing Facet				
PoS information	X	X	X	X

A basic parameter of navigation is the manipulation of maps, the use of menus and the interchange between the classic app user interface and the map's user interface. Other mechanisms such as indexes and appropriate dialogs help the user navigate anywhere in the application. Also embedded search engines provide the ability to find information by entering keywords or parameter queries using logical operators. Important parameters in the search and processing of information are the correlation and relevance of the data

retrieved in relation to the query and the response rate of the application to user requests.

The routing process is based on the location of the user and on the ability to provide routes to POI or PoS around it based on his/her preferences and needs. The information process includes the mechanisms by which the user retrieves information about entities of interest, without necessarily referring to destination of a route, but also information about a situation such as a traffic jam.

TABLE IV. 2ND LEVEL MAPPING FOR FUNCTIONALITY

Functionality				
Level 2	Sub-characteristic			
	Suitability	Precision	Interoperability	Security
	Correlation r_{ij}			
Presentation Facet				
Maps -3D road view	X	X	X	O
Maps-Satellite view	X	X	X	X
Maps - traffic information	O	O	X	O
Maps - friends close by	X	O	O	O
User status info (speed, direction, coordinates)	X	X	X	O
Navigation Facet				
Voice command	X	X	O	O
Change map view angle	X	X	O	O
Routing Facet				
Routing to friend position (moving POI)	X	O	X	X
Re-routing (user on the move)	X	X	O	O
Information Facet				
Update moving POIs position	X	X	X	O
Update real-time traffic conditions	O	O	X	O
Update real-time traffic events	O	O	X	O
Purchasing Facet				
Purchase	X	X	X	X

TABLE V. 3RD LEVEL MAPPING FOR FUNCTIONALITY

Functionality				
Level 3	Sub-characteristic			
	Suitability	Precision	Interoperability	Security
	Correlation r_{ij}			
Presentation Facet				
Maps -3D road view	X	X	X	O
Maps-Satellite view	X	X	X	X
Maps - traffic information	O	O	X	O
Maps - friends close by	X	O	O	O
User status info (speed, direction, coordinates)	X	X	X	O
Navigation Facet				
Voice command	X	X	O	O
Change map view angle	X	X	O	O
Routing Facet				
Routing to friend position (moving POI)	X	O	X	X
Re-routing (user on the move)	X	X	O	O
Information Facet				
Update moving POIs position	X	X	X	O
Update real-time traffic conditions	O	O	X	O
Update real-time traffic events	O	O	X	O
Purchasing Facet				
Purchase	X	X	X	X

D. Mathematical Model

Let π_x be a quality characteristic of ISO9126, with $x=1,2,3,4$. Thus, π_j is the Functionality quality characteristic of ISO9126. Let $\lambda(\pi_x)$ be the number of quality sub-characteristics of quality characteristic π_x , e.g. from the ISO1926 definition [4], it holds that $\lambda(\pi_1)=4$.

IV. EXPERIMENT

Let $\beta_{\Pi_{x,j}}$ denote the significance weight of a sub-characteristic (SCSW) taking values in the interval [0,1], for the quality characteristic π_x and for its quality sub-characteristic j ($j=1.. \lambda(\pi_x)$).

It holds that:

$$\pi_x \sum_{j=1}^{\lambda(\pi_x)} \beta_{\pi_{x,j}} = 1 \quad (2)$$

This means that the sum of the SCSWs for each qualitative characteristic equals to 1, e.g. for Functionality that includes four sub-characteristics, it holds:

$$\beta_{1,1} + \beta_{1,2} + \beta_{1,3} + \beta_{1,4} = 1 \quad (3)$$

The same holds for β_{Π_x} which denotes the significance weight of a characteristic (CSW).

Let F_μ denote the facet ($\mu=1..5$) and L_i the level of significance ($i=1..3$). Let W_{F_μ, L_i} (taking values in the interval [0..1]) denote the Total Significance Weight (TSW) for facet F_μ and significance level L_i . Then it holds that for each quality characteristic, the sum of all TSW equals to 1, for each facet, e.g. it holds that:

$$W_{F_1, L_1} + W_{F_1, L_2} + W_{F_1, L_3} = 1 \quad (4)$$

Where F_1 is the Presentation Facet and L_1-L_3 the three levels of significance.

Using these definitions, the quality assessment model defines significance weights for characteristics (CWS) and sub-characteristics (SCWS), which express the emphasis that needs to be given during system analysis and design. The value of the weight of each feature depends on the emphasis system designers wish to give to a system based on expert opinion, since this is more of a strategic decision. For example, a strategic design decision would be to emphasize more on suitability than on precision. Furthermore, it defines significance weights for facets per significance level (TSW). This weight denotes the emphasis designers wish to give on basic functions of the system (significance level 1) or advanced functions (levels 2 and 3). Users determine these values directly.

The correlation function takes a numeric value (SW) based on the results of the experiment. Let σ_β denote the SW of component/function (σ). Then, for the interaction facet F_μ and for the significance level L_i that corresponds to the quality sub-characteristic $\lambda(\pi_x)$ of the quality characteristic π_x , SW is calculated by the formula:

$$\sigma_\beta = \sum_{j=1}^{\lambda(\pi_x)} r_{ij} \beta_{\Pi_{x,j}} \quad (5)$$

The normalized values of SW, $\kappa\sigma_\beta$ take values in the interval [0,1].

At the quality characteristic level, the Composite Significance Weight (CSW) w_σ is used, that is the combined SWs of each function per significance level per facet per quality characteristic:

$$w_\sigma = \kappa\sigma_\beta * \beta_{\pi_x} * W_{F_\mu, L_i} \quad (6)$$

A. Experiment Setup

Following the first stage of the development of the model, which included the determination (by an expert on quality) of the correlation function between the system components and the four external qualitative sub-characteristics of ISO9126, the second step details the value of this correlation (strength of relation). The Composite Significance Weight (CSW) for each function-quality characteristic relation is defined as the qualitative value of the structural elements of the model as given by normalized numerical values in the interval [0,1]. The values of SW were calculated using two methods a) through the judgment / opinion of an evaluator; and b) through an experiment involving users executing predefined scenarios. User data were collected using a structured questionnaire. The values set by the evaluator and the resulting values from the completion of the user responses were combined using the Quality Model's mathematic formula to extract the final CSW.

The mobile application used in the experiment was the Google Maps app, an app that is considered both popular and user-friendly. A variety of platforms and operating systems was used including smartphones equipped with the Windows Mobile operating system or the Android OS, GPS receiver, touchscreen and wireless 4G data transfer protocols. The user sample surveyed included 5 experienced users who had used at least 10 times the specific or similar navigation applications. Users were asked to perform 12 specific multi-step scenarios in a predefined way, evaluating the quality of the components of the quality model when interacting with the system. The tests were designed to include only the functions/components of the app implemented in Greece, excluding some functions such as real-time traffic update (which, although supported by the Greek version of the app, real time updating is not supported) or in app purchases using PoS. Following the scenario enactment, users completed a structured questionnaire (the Likert type rating scale was used), evaluating the system's operation in real conditions. Users evaluated all ISO9126 characteristics and sub-characteristics for all LBS functions/components detailed in Section II. Evaluation was organised per facet and per level of significance. Correlations that were not recognised were evaluated with '0' and with a '-' (dash) if the function/component was not included in the performed scenario.

B. Calculating and Assigning Values

The correlation table was initially defined by the evaluator before user participation. However, from the processing of the user questionnaires, differences in estimates were observed for some functions/components leading to a slightly updated version of the table. In the current experiment, there was a chance that few discrepancies and/or inaccurate responses may influence overall results so the parameter values were categorised into two evaluation clusters. The first cluster of values was formed based on the expert estimates and the other on user responses. Normalized significance weights were calculated by using the table of values of significance weights for each qualitative sub-characteristic resulting from the first two processing steps, as well as the correlation tables of users and evaluator. The resulting significance weights were

calculated separately based on both the user association table and the evaluator table. If no operation was performed by any user, then a dash ('-') was the corresponding weight value for that component / function. The calculation of the composite weight of significance was performed by taking into account the significance weights of each qualitative characteristic (initially a weight of 0.25 / characteristic was assigned) and the weights of each significance level (initially a weight of 0.6, 0.3 and 0.1 was assigned to each of the 1st, 2nd and 3rd levels, respectively). This process was based on the correlation tables of both the users and the evaluator. If a component had not been assigned a value, then it was not evaluated and the corresponding cell was filled-in with a dash ('-').

C. Experiment Results

The resulting final tables depict the app functions/components with assigned CSW values, sorted by level, appearance and quality characteristic, for each of the three significance levels (depicted in Tables VI-VIII for significance levels 1, 2 and 3, respectively). These tables present values assigned by the users.

TABLE VI. 1ST LEVEL COMPOSITE SIGNIFICANCE WEIGHTS VALUES

ISO9126				
Level 1	Characteristics			
	Functionality	Usability	Performance	Reliability
Composite Significance Weight (Wσ)				
Presentation Facet				
Maps (routing)	0,038	0,0375	0,050	0,038
Maps (user position)	0,038	0,0375	0,050	0,038
Text info (info tags)	0,037	0,0375	-	0,038
Navigation Facet				
Move within map	0,025	0,025	0,038	0,026
Focus	0,025	0,025	-	0,026
2D presentation	0,025	0,025	0,038	0,026
Dialog Screens	0,025	0,025	0,038	0,026
Help	0,025	0,025	-	0,019
Routing Facet				
My position	0,041	-	0,050	0,050
Search for POI/PoS	0,041	0,075	0,050	0,050
Show route(s)	0,041	0,075	0,050	0,050
Information Facet				
POI information	0,15	0,15	0,15	0,15
Purchasing Facet				
PoS information	0,10	0,10	0,10	0,10

TABLE VII. 2ND LEVEL COMPOSITE SIGNIFICANCE WEIGHTS VALUES

ISO9126				
Level 2	Characteristics			
	Functionality	Usability	Performance	Reliability
Composite Significance Weight (Wσ)				
Presentation Facet				
Maps (POIs)	0,0226	0,019	0,019	0,019
Maps (Object location)	0,0226	0,019	0,019	0,019
Text info (POI/user info tags)	0,0220	0,019	0,019	0,019
Navigation Facet				
Focus	0,015	0,015	0,025	0,015
Hide/Show POIs	0,015	0,015	-	0,015
Analyse POI	0,015	0,015	0,025	0,015
Dialog Input Screens	0,015	0,015	0,025	0,015
Shortcuts	0,015	0,015	-	0,015
Routing Facet				
History	0,019	0,025	0,019	0,015
Favourites	0,019	-	0,019	0,020
Mode (Pedestrian, Car etc.)	0,018	0,025	0,019	0,020
Reverse	0,018	0,025	0,019	0,020
Information Facet				
POI position	0,038	0,038	0,038	0,038
POI detailed information	0,038	0,038	0,038	0,038
Purchasing Facet				
PoS detailed information	0,038	0,038	0,038	0,038

Results of this particular experiment demonstrate the fact that the evaluation based on both the users and the evaluator, in defining the correlation function, are largely yielded almost the same ordering of system components/function (Table IX). As far as Functionality is concerned, the most basic functions/components were highly rated thus analysts, designers and engineers should attach great importance to their quality analysis and design. Considering basic user requirements, locating ones position searching and managing POIs and creating alternative routes were positively evaluated. Excluding usability, there were no significant differences between the results from the users' and evaluator's correlation estimates. The most valuable functions/component were the

indication of POIs, while followed by the presentations of the route through maps for users and POI search. The most efficient and reliable function/components were the provision of information on points of interest. Therefore, the most qualitative components are those that serve the basic functions expected to support a navigation application.

The overall conclusion from this experiment was that the specific application provides all the necessary functions for the large majority of users, and is deemed reliable (except from cases where the position of the user was not pinpointed with the same accuracy, re-routing algorithms took much longer time to calculate alternatives than anticipated and real-time data was not available). Usability seemed to be a concern especially for users with small screen devices where information overload seemed to be a problem, especially when moving. Users also rely more and more on additional information for POI/PoS, especially on other peoples' opinion and ratings. A careful interpretation of the results produces user requirement categorised by a formal qualitative perspective and thus helps designers of an existing app to develop a better new version.

TABLE VIII. 3RD LEVEL COMPOSITE SIGNIFICANCE WEIGHTS VALUES

ISO9126				
Level 3	Characteristics			
	Functionality	Usability	Performance	Reliability
Composite Significance Weight ($W\sigma$)				
Presentation Facet				
Maps -3D road view	0,038	0,038	0,019	0,019
Maps-Satellite view	0,06	0,008	0,008	0,008
Maps - traffic information	0,04	0,008	0,008	0,006
Maps - friends close by	0,06	0,008	0,008	0,006
User status info (speed, direction, coordinates)	0,005	-	-	-
Navigation Facet				
Change map view angle	0,012	0,013	0,013	0,013
Routing Facet				
Routing to friend position	0,008	0,010	0,013	0,013
Re-routing (user on the move)	0,009	0,010	0,013	0,013
Information Facet				
Update moving POIs position	0,009	0,013	0,017	0,013

V. DISCUSSION AND CONCLUSIONS

Systems that use location based sub-systems combine software applications, hardware and networks to provide a high level of interaction with the user. LBS-enabled services are aimed at a broad spectrum of mobile users and thus, the capture, organisation, classification and satisfaction of user requirements during the analysis and (re-)design phases are a challenge.

Users interacting with such systems seek ease of use, fast responses, autonomy, financial gain, enjoyable navigation experience tailored to their personal needs. The degree to which user requirements are satisfied affects the success of a system and characterize its quality of use. The quality of LBS-enabled systems can be evaluated against quality of applications that support the system and quality services provided by the system. Evaluating systems based on quality is a means to derive (ever-changing) user requirements than may, in turn, be used to re-design a system or design a new one. To this end, evaluation can be approached using two complementary perspectives: the evaluation of the functions supported by the software and the evaluation of the services provided to the user. The evaluation of software functions requires specialized knowledge and can be performed by software engineers who can also act as evaluators. Experts are able to contribute to the hierarchical analysis of quality from general to partial. However, the user of experts will not suffice, software must also be evaluated by the final users during its use.

TABLE IX. THE TOP-3 QUALITY FUNCTIONS PER FACET

Facet	Top 3	Function/Component	
		Correlation (Users)	Correlation (Experts)
Presentation	1	Map (routing)	Map (routing)
	2	Map (POI)	Map (POI)
	3	Map (road view)	Text info (POI/user info tags)
Navigation	1	Move within map	Move within map
	2	Focus	Hide/Show POIs
	3	Change map view angle	Change map view angle
Routing	1	Search for POI/PoS	Search for POI/PoS
	2	Reverse	Reverse
	3	Favourites	Mode (Pedestrian, Car etc.)
Information	1	POI position	POI position
	2	POI information	POI information
	3	Update moving POIs position	POI detailed information

In this paper we presented a Quality Model that identifies and analyzes the basic components of location-based enabled e-commerce software, and determines their correlation to ISO 9126 quality features. The model's goal is to detailed analysis of the quality of the user's requirements specifically for the LBS sub-system. It analyzes the system into components/functions, which are in turn categorised into aspects of interaction (facets) with the user and levels of significance. The next step involves the mapping of the components to the four external quality sub-characteristics of the ISO9126 standard for quality evaluation, via the definition of a suitable correlation function. The mathematical foundation of the model permits the calculation of values for these weights that depict the importance of specific system features to the designers or to the users. Goal-oriented design is supported when evaluators set quality targets (weights) to facets, levels of significance and or quality characteristics. Users express their requirements by setting (through evaluation) the weights for functions and sub-quality characteristics. The use of formal standards for the evaluation of functions/components also enables the use of a common vocabulary across analysis, design and evaluation teams.

REFERENCES

- [1] N. Ahmad, A. Rextin, U.E. Kulsoom, "Perspectives on usability guidelines for smartphone applications: An empirical investigation and systematic literature review," *Information and Software Technology*, vol. 94, pp. 130-149, 2018.
- [2] P. Heck, A. Zaidman, "A systematic literature review on quality criteria for agile requirements specifications," *Software Quality Journal*, vol. 26(1), pp. 127-160, 2018.
- [3] Q. Zheng, Y. Han, S. Li, J. Dong, L. Yan, J. Qin, "E-commerce Architecture and System Design," in *Introduction to E-commerce*, Q. Zheng Q., Eds. Berlin: Springer, 2009.
- [4] ISO/IEC 9126: 2001, "Software engineering -- Product quality -- Part 1: Quality model". Zyrich: ISO, 2001.
- [5] ISO/IEC 25010: 2011 "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuARE) — System and software quality models", Zyrich: ISO, 2011.
- [6] N.A. Ernst, A. Borgida, I.J. Jureta, J. Mylopoulos, "An overview of requirements evolution," in T. Mens, A. Serebrenik, A. Cleve eds. *Evolving software systems*. Berlin: Springer, pp. 3-32, 2014.
- [7] S. Dhar, U. Varshney, "Challenges and business models for mobile location-based services and advertising," *Commun. ACM*, vol. 54(5), pp. 121-128, 2011.
- [8] H.P. Breivold, I. Crnkovic, "Analysis of Software Evolvability in Quality Models," *Software Engineering and Advanced Applications 2009, SEAA '09, 35th Euromicro Conference on*, pp. 279-282, 2009, ISSN 1089-6503.
- [9] W. Alsaqaf, M. Daneva, and R. Wieringa, "Quality Requirements in Large-Scale Distributed Agile Projects - A Systematic Literature Review," in *REFSQ2017, 2017*, vol. 10153, pp. 219-234.
- [10] E. Parra, J.L. de la Vara, L. Alonso, "Analysis of requirements quality evolution," in *Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings (ICSE '18)*, 2018, ACM, New York, NY, USA, pp. 199-200.
- [11] A. Stefani, M.N. Xenos, "E-commerce system quality assessment using a model based on ISO 9126 and Belief Networks," *Software Quality Journal*, vol 16(1), pp. 107-129, 2008.
- [12] J.D. Garofalakis, A. Stefani, V. Stefanis, "A Framework for the Quality Evaluation of B2C M-Commerce Services," *IJHCR*, vol. 2(3), pp. 73-91, 2011.
- [13] H.T. Kanwal, F. Arif, A.M. Zaidi, "Software requirement engineering a new leave towards the silver bullet," *Science and Information Conference (SAI)*, pp. 189-198, 2015.
- [14] D. Granlund, D. Johansson, K. Andersson, R. Brännström, "A Case Study of Application Development for Mobile and Location-Based Services," in *Proceedings of International Conference on Information Integration and Web-based Applications & Services (IIWAS '13)*, New York: ACM, 2013.
- [15] S. Ilarri, E. Mena, A. Illarramendi, "Location-dependent query processing: Where we are and where we are heading," *ACM Comput. Surv.*, vol 42(3), pp. 1-73, 2010.
- [16] ISO/IEC 42010: 2011, "Software engineering -- Product quality -- Part 1: Quality model Systems and software engineering -- Architecture description". Zyrich: ISO, 2011.
- [17] J.M.S. França, M.S. Soares, "SOAQM: Quality Model for SOA Applications based on ISO 25010," in *Proceedings of the 17th International Conference on Enterprise Information Systems*, vol 2 (ICEIS 2015), Slimane Hammoudi, Leszek Maciaszek, and Ernest Teniente eds.), Vol. 2. SCITEPRESS - Science and Technology Publications, Lda, Portugal, pp. 60-70, 2015.

Programming Technologies for the Development of Web-Based Platform for Digital Psychological Tools

Evgeny Nikulchev¹, Dmitry Ilin²
MIREA – Russian Technological
University & Russian Academy
Science, Moscow, Russia

Pavel Kolyasnikov³
Vladimir Belov⁴
Russian Academy Science
Moscow, Russia

Ilya Zakharov⁵, Sergey Malykh⁶
Psychological Institute of Russian
Academy of Education
Moscow, Russia

Abstract—The choice of the tools and programming technologies for information systems creation is relevant. For every projected system, it is necessary to define a number of criteria for development environment, used libraries and technologies. The paper describes the choice of technological solutions using the example of the developed web-based platform of the Russian Academy of Education. This platform is used to provide information support for the activities of psychologists in their research (including population and longitudinal researches). There are following system features: large scale and significant amount of developing time that needs implementation and ensuring the guaranteed computing reliability of a wide range of digital tools used in psychological research; ensuring functioning in different environments when conducting mass research in schools that have different characteristics of computing resources and communication channels; possibility of services scaling; security and privacy of data; use of technologies and programming tools that would ensure the compatibility and conversion of data with other tools of psychological research processing. Some criteria were introduced for the developed system. These criteria take into account the feature of the functioning and life cycle of the software. A specific example shows the selection of appropriate technological solutions.

Keywords—Psychological research tools; web-based platform; choice of the tools and programming technologies

I. INTRODUCTION

Currently, computer technologies are actively used for data collection in the field of education. In recent years, web-based technologies are going to be widely known for psychological researches. Computers are used not only for questionnaires automation, but also for complex cognitive tests, that contains different graphics. One of the first attempts for cognitive tests automation was undertaken in the TAPAC (Totally Automated Psychological Assessment Console [1]) system that consisted of a console for test subject's answers, tape recoder and projector. Since then, a number of technological resources for psychological testing automation have increased. Internet technologies development radically changed possibility for data collection. First, they make it possible to increase the data set of research. Secondly, web-based technologies reduce the time of data collection and the cost of studies [2]. The development of modern computing technologies provides new opportunities for organizing large-scale population studies of the psychological characteristics of students. The results of these studies can be used for the national standardization of psychometric tools.

In addition, large accumulated data sets can become the basis for machine learning mechanisms and other approaches using artificial intelligence. Accumulation of data from population studies into a single system can allow a breakthrough in the development of systems for automated intellectual analysis of behavior data.

The issue of selecting methodological tools for online and offline research includes several items.

First, any selection presupposes the existence of generally well-defined criteria, on the basis of which a decision can be made to include or not to include techniques in the final toolkit.

Secondly, an important factor in the ongoing research is the separation of them into online and offline methods. Creation of new research tools based on web technologies will allow creating not only complex experimental psychological models, but also increase the statistical power of the received data due to the expansion of research samples. The choice of methods based on this division, on the one hand, imposes certain restrictions on the selection of tools, on the other hand, allows to focus on the strengths and weaknesses of the methods used.

The selection of methodological tools in the field of education and psychology has a number of common for methods of studying the behavior of criteria (requirements for reliability and validity of methods, etc.), and specific parameters (related to the field of application). Common criteria for the selection of tools include generally accepted requirements for the reliability of psychological testing tools. These include:

- Assessment of the quality of testing, which includes primarily an analysis of the knowledge, skills, abilities or individual characteristics that need to be assessed; quality assessment involves the construction of a specific goal and test criteria.
- Validity of the tool (validity and suitability of application of methods and results of research in specific conditions).
- The reliability of the tool (the possibility of obtaining identical results in subjects in different cases).

- The reasonableness of the methodology, the presence of an adequate psychological theory, which underlies the methodology.
- The conformity of the methodological tool to the cultural norms of the estimated population. This is one of the important indicators, which is often underestimated, assumes the account of the so-called "bayes" or hidden context, understood by the testing participant in the answers to the questions of the methods, depending on the cultural norms. This includes the equality of all participants in testing, the absence of harassment on the basis of gender, national, religious grounds, any issues that may somehow hurt feelings or adversely affect a participant.
- The existence of an explicit empirical mechanism for interpreting assessments of the subject, such as population norms, criteria for determining clinical groups and criteria for classifying individuals.
- The quality of administrative, interpretive and technical guidelines. The presence of a clear procedure for the conduct and interpretation of individual techniques, as well as batteries in general. Uniformity in the interpretation of the results obtained and their use.
- The basis for conclusions about broader underlying behaviors and attributes from the pattern of behavior.
- Ease of use of the test material.

Amazon Mechanical Turk platform was the first widely used web-based psychological tool. This platform allowed researchers for a small amount of time to hire individual participants to fill out individual psychological techniques online. MTurk was used mainly by researchers from the United States. It was shown that MTurk samples do not fully correspond to the characteristics of the US population. For example, the population of MTurk was mostly white and female, and also more educated and younger than the US population as a whole. However, the quality of the data obtained remained fairly good [3]. MTurk was successfully used to study the attention [4], creativity, dishonest behavior [5] and sexual attitudes [6]. It has been shown that MTurk can be a valuable tool even for working with a clinical population [7]. To date, the Russian-language Yandex.Tolok platform, similar to the MTurk platform, has been developed. After the success of MTurk, new technologies began to appear, but the approach has changed. While MTurk was an online marketplace with an audience of its own, new tools provided only tools for creating research, which could then be distributed over the Internet, while the researcher himself had to provide a sample. Google Forms service (Google Inc.) was one of the most famous tools of this type. This free software was designed to create web surveys. Every researcher could create his own questionnaire, for which he was given a unique link that he could distribute on the Internet. Google Forms also provided simple analytics for researchers. Its openness and brand Google made it ubiquitous for online surveys, especially among students. However, Google Forms does not have a number of functions that play an important role in psychological research. For example, an analysis of the

response time characteristics based on the Google Form Service is not possible. There are also problems with storing the collected data. In this regard, for psychological needs, additional products have been developed. Among the most recognized we can mention Survey Monkey, Qualtrics, LimeSurvey or EnKlikAnketa. Some of these products are paid, or at least shareware (shareware: SurveyMonkey, Qualtrics), some of them are free (EnKlikAnketa). These services can provide many useful functions for researchers. For example, the Slovenian service EnKlikAnketa offers assistance in correcting methodological deficiencies in the development of surveys. It also allows to collect a lot of metadata, such as the site from which the respondent went to fill out the survey, the time spent on the poll, or the characteristics of the browser and operating system of the computer on which the respondent was working.

To date, this type of services has almost replaced traditional research using paper forms. However, they can only be used for questionnaires or knowledge tests (q-type data according to Cattell [8]), whereas studies in the field, for example, of individual differences in cognitive or control functions require a wider functional. In these areas, a class of modern computerized technologies is being developed. Computerized presentation of tests is convenient because of several factors. First, it helps to automatically control the process of presenting tasks, thereby reducing errors related to the human factor. Secondly, computerized tools of this type can record aspects of performance with accuracy not available for other methods. Among such characteristics of the tests, one can single out an estimate of the reaction time, an estimate of the exact spatial position of the cursor on the screen, and others. Various cognitive tests are often organized into batteries that are designed for a comprehensive assessment of the cognitive domain. Widely used cognitive characteristics include general cognitive abilities, working memory [9], spatial or mathematical abilities [10]. Most of these batteries have been developed for clinical use NAB [11]. However, they have also been used successfully to study regulatory samples [12], as well as for research in the field of behavioral genetics [13].

The computerized application of cognitive tests is becoming more accessible to researchers in the field of psychology. First, there are already free software solutions, such as PsychoPy (<http://www.psychopy.org/>), which allow psychologists to develop their own tests in the absence of advanced programming skills. Some of the applications of this type even contain their own battery tests. For example, there is a programming language for creating tests for psychology (PEBL, <http://pebl.sourceforge.net/>). It requires more skills to create a specific test than PsychoPy, however, a large set of pre-programmed executables that can measure a wide range of characteristics is freely available to it.

At present, the main limitation of automatic batteries is their possibility to be controlled only autonomously with the help of downloaded and pre-installed software. The psychological community tends not to trust the accuracy of the tests conducted online, due to potential side variables, such as the technical properties of personal computers or respondents' monitors. However, the quantification of technical noise

shows that at least for some types of tasks, web experiments can be an acceptable source of data [14]. Thus, the next logical step for computerized cognitive tests is to distribute via the Internet in the same way that it happened with the questionnaires. In accordance with this, the PsychoPy team, for example, recently presented the possibility of launching experiments in a web browser.

Summarizing, it is possible to formulate the main advantages of computerized and web technologies for research in the field of psychology. The main advantages are:

- accessibility for large-scale research;
- the increase in reliability and the potential for generalization of the results obtained;
- lower costs for equipment of premises for experiments;
- the opportunity to avoid all the troubles associated with the use of laboratories: (booking, limited space, the need for expensive specialized equipment, etc.);
- the ability to provide tools around the clock without any time limits;
- the possibility of open research, with a fully voluntary participation, which usually improves the motivation of respondents.

It is also necessary to remember the potential difficulties associated with this type of research. Most of the difficulties can be related to technical control over the comparability of tests conducted on computers with different system characteristics. In addition, the researcher's control over the progress of the test is reduced (the research participant can perform tasks alone without additional supervision).

At present, many tools are used for software development. These tools differ in their functionalities and programming convenience, as well as they are not without disadvantages, that often appear only in development stage, when system extension or modules integrating. Therefore, it is an important task to choose toolset and programming technologies in the planning stage. This choice should satisfy the software requirements and programming process. In this case, it is necessary to consider the parameters of the technologies [15], the guaranteed quality of data processing [16], reliability with extension [17] etc.

The aim of the paper is to describe the choice process of the technological solutions on the example of the being developed digital web-based platform. The platform provides information support for psychologists' activities in conducting research (including population and longitudinal researches) [12, 18].

Software architecture is not only a structural basis for system components and their connections describing, but also determines the approaches to development and environment. The architecture description should include answers to the questions that arose during the system designing. Described service is web-based platform that includes server and client sides. Therefore, one of the main tasks is to choose the programming language and technologies that are most suitable

for components development [19]. The platform should work in most browsers, including mobile devices, without installing any plug-ins or extensions. Therefore, it is necessary to choose the solutions that will not make any certain restrictions or need installing additional plug-ins and libraries. The requirements for server-side components are less restrictive. However, it is necessary to take into account the features of the technological solutions and the complexity of the result software.

Therefore, there are following system features:

- significant amount of developing time that needs implementation and ensuring the guaranteed computing reliability of a wide range of digital tools used in psychological research;
- ensuring functioning in different environments when conducting mass research in schools that have different characteristics of computing resources and communication channels;
- possibility of services scaling;
- security and privacy of data;
- use of technologies and programming tools that would ensure the compatibility and conversion of data with other tools of psychological research processing.

To achieve effective solution a number of techniques were used. The primary stage included architecture requirement analysis. The aim of this stage is to identify the main use case, functional and non-functional requirements for the web-based platform [20].

Based on the received information, architecture synthesis was carried out. The reason is to determine a set of coupled components of the system, their connections, the most effective ways of data exchanging. To choose concrete programming language and technologies usable for web-based platform development their study and comparing were carried out. It was completed in the context of the formed architecture, existing requirements and constraints. As for programming languages for browser applications, the ability of application delivering without the need for installation additional software was evaluated. Frameworks are considered for their active application in projects, their community and relevance of the task. It is worth to note that the direct comparison of frameworks will not give concrete result. However, a number of these frameworks are more suitable due to better scalability, less costs for study and more ready-made modules.

The paper includes following ensuing sections. The II and III parts describe the features of the being developed system architecture and its main components. In the IV part, criteria are introduced for the client and server sides of the application. The section V describes the choosing of the appropriate technological solutions.

II. PLATFORM ARCHITECTURE DESCRIPTION

For the formation of an adequate architectural solution, a number of methods have been applied. The initial phase included the architectural requirement analysis, in order to

identify the main uses, functional and non-functional requirements for the platform [20]. In addition, to clarify the requirements, unstructured interviews of the pedagogical staff were used to reveal the degree of variation in the technical characteristics of software and hardware in Russian schools.

From the perspective of the end-user, the project will consist of two main components: the researcher's private account (Fig. 1) and offline applications for offline experiments (Fig. 2).

Given the fact that the number of users will grow over time, the web service should be scalable horizontally (Fig. 3). Every API node should include a multi-level architecture. In combination with Object-DocumentMapper (ODM), it will provide more flexibility than monolithic architecture.

Experimental and intensive data algorithms for demographic research should be separated from the main service, as well as from administrative functions (Fig. 4). From a security perspective, the administration panel can be used as a separate service on the intranet.

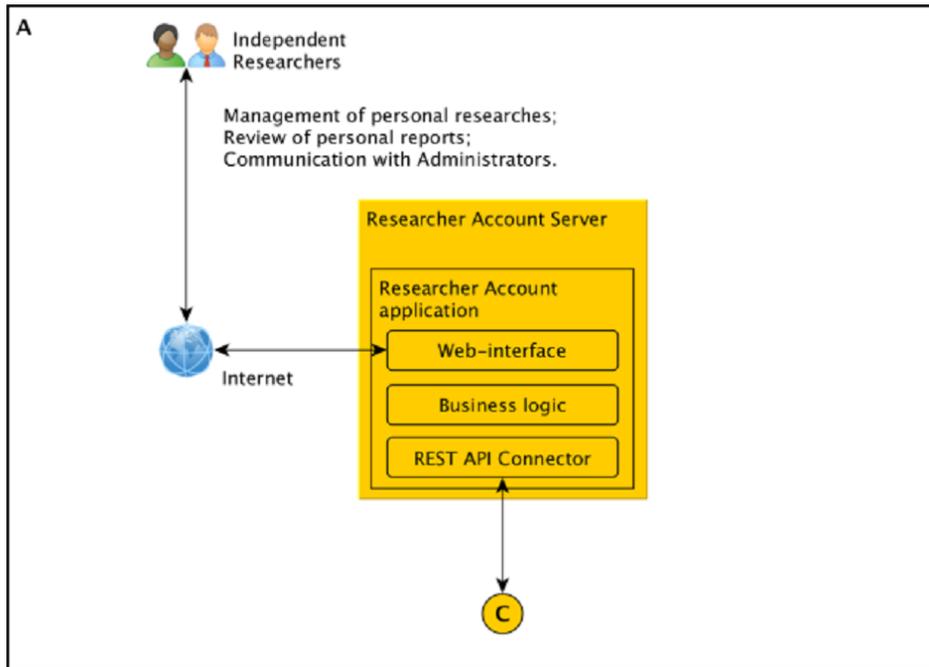


Fig. 1. An Independent Researcher Tool.

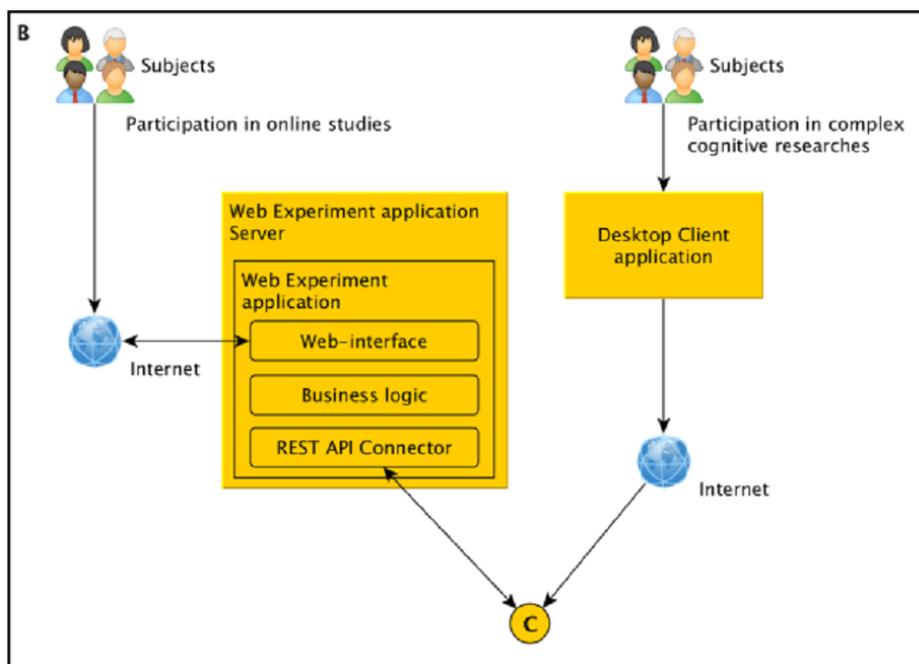


Fig. 2. Tools for Online and Offline Experiments.

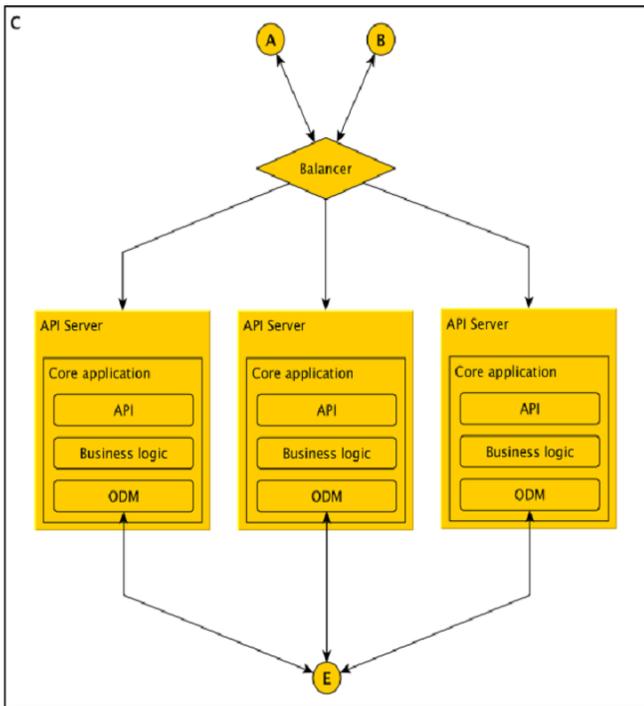


Fig. 3. Scalable API Server for Unified Data Access and Management.

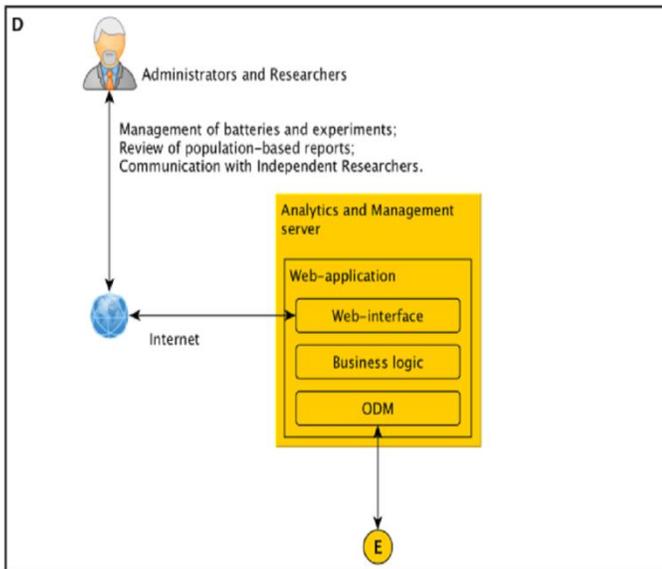


Fig. 4. Isolated Analytical Tools for Population Studies.

In addition to scaling the algorithmic part, data storage should also be scalable [21, 22]. The best approach for the project is the combination of sharding and replication (Fig. 5). Sharding allows to distribute data between different physical servers (shards) based on the value of some key, so that the entities are grouped into data set for this key. Replication allows to copy data between several servers, among which one server (master) for data saving, and others (slave) - for reading. Thus, sharding can provide a system with high I/O performance, while replication can help to ensure the availability of the service.

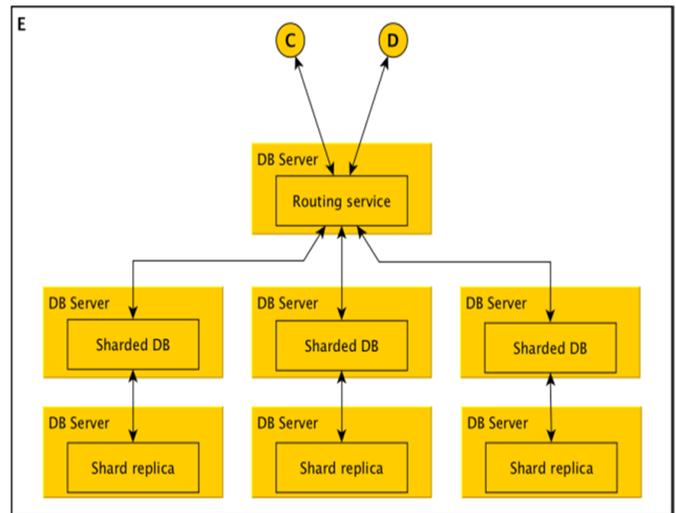


Fig. 5. Horizontal Scaling of Storage by Means of Sharding and Replication.

III. DESCRIPTION OF PLATFORM COMPONENTS

The architecture of the being developed platform for psychological research was chosen to be multi-component, which provides more flexibility than monolithic.

Monolithic architectures have a number of disadvantages:

- the larger the system, the more difficult it is to maintain it and make changes;
- with a large system, changing a small portion of the code can cause errors in the entire system;
- after each code change, it is necessary to test the entire system for errors.

Unlike monolithic, the use of multi-component architecture gives the following advantages:

- writing and maintaining smaller parts is easier than one large system;
- it is easier to distribute the developers to write a specific part of the system;
- the system can be heterogeneous, because for each component it is possible to use own languages and technologies, depending on the task;
- easier to upgrade because only the required component is affected;
- the system becomes more fault-tolerant, since in the event of failure of one and the components, others may still be working.

Thus, the choice in favor of a multi-component architecture is justified in view of a number of advantages over the monolithic and the most suitable taking into account the requirements for the developed platform.

Fig. 6 shows the scheme of the platform architecture for psychological research. The architecture is divided into separate components that can work independently and communicate among themselves using the REST API.

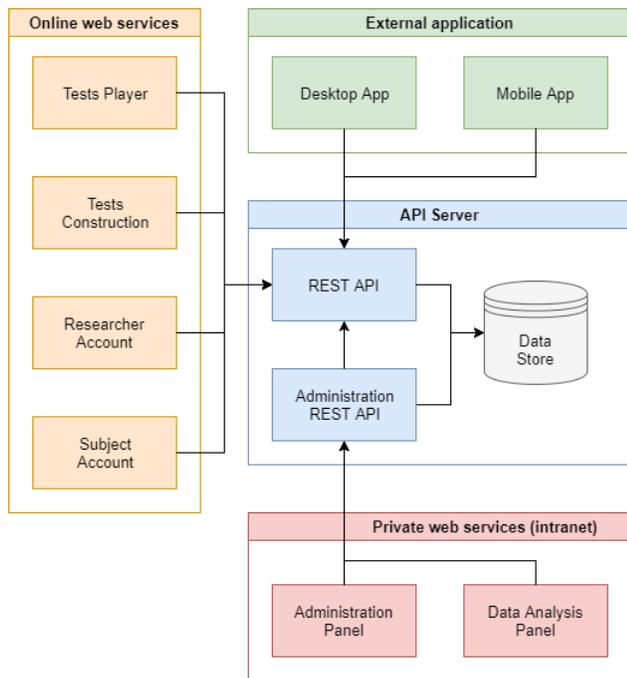


Fig. 6. Schematic of a Multi-Component Platform Architecture.

«**APIServer**» is the main core of the system, which is a RESTAPI server and is responsible for working with the data store, as well as for performing various service functions.

«**Onlinewebservices**» are components that should be accessible from the Internet. They are the main online part of the platform being developed, among which are the online test player, the online test designer, the researcher's personal area and the personal area of the examinee.

«**Externalapplication**» is separate application, such as desktop and mobile. Unlike the online version of the test player, the feature of the applications is that the process of passing these tests should work without connecting to the Internet. In this regard, the data for the tests should be loaded in advance, and after passing the tests the subject is uploaded back to the server.

«**Privatewebservices (intranet)**» are separate services, including the Platform Administration Panel and the Population Analysis Toolbar. The feature of these services is their need to be isolated from direct access from the Internet for a greater security guarantee. It is also worth noting that these services communicate with their own RESTAPI, which includes administrative methods, which should not be accessible from the Internet.

To carry out psychological research, many tests are used, which can include necessary materials, for example, images or files.

Thus, for storage and transmission of tests, a batch approach is relevant, which will allow storing and transferring data in one file. The use of the package is also important for transferring data to the client part of the application, where it cannot always be guaranteed access to the Internet. It includes both an online application and a desktop application that will

be used in schools, as well as an application for mobile devices. It is possible to draw an analogy of the test suite with data for city maps, which are downloaded as a package and then unpacked onto the device.

Packages will be used both for sending psychological tests to the application and for obtaining test results and sending them to the server.

Using a package for data transferring to online applications in the client browser is justified by following reasons:

- one request to the server is used, instead of several, which minimizes the load on creating an HTTP connection;
- the client side does not depend on the server at the time of passing the test;
- in the case of Internet disconnection, the user can complete his research;
- logging time on the client side minimizes the errors of the record;
- it is easier to track the degree of workload of the package and to inform the user about it.

Fig. 7 shows the structure of the package for storing the tests. It shows that the package is complete, includes information and description of this package, as well as the tests themselves. There should be at least one test in the package. The test includes data on its description, as well as images and files, if they are needed.

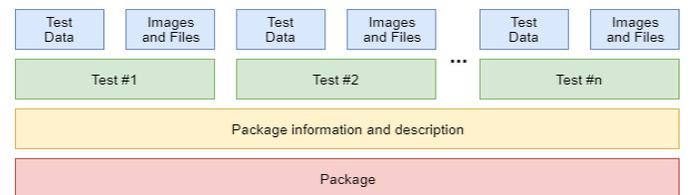


Fig. 7. Diagram of the Structure of the Package for Storing and Transferring Tests.

The test should be described using a special JSON Schema standard, the structure of which is approved in advance. Based on this structure, the test will be validated.

Using JSON Schema avoids a number of problems and has the following advantages:

- no need to manually check the contents of the documents;
- no need to create own validators with a variety of configurations and takes care of the support of these decisions;
- due to a single standard, the process of integration and support of validation in various components of the platform is simplified, such as online test player and desktop applications;
- the change in the scheme does not require the replacement of the validator code;

- it is possible to describe the psychological test manually without the help of additional tools, which will be a plus at an early stage of development, when the test designer is not yet available;
- there are a large number of implementations for various programming languages and platforms.

Based on the requirements to the platform for digital psychological research, a list of virtual machines required for their deployment on the server was generated (Table I).

TABLE I. VIRTUAL MACHINES

Purpose of the VM	Number of VM (pcs)	OS Version	RAM (GB)	CPU (pcs)
API server	2	CentOS 7.3	8	4
Personal cabinet of a psychologist	1	CentOS 7.3	2	1
Online survey system	1	CentOS 7.3	2	1
Internet portal	1	CentOS 7.3	2	1
Intranet administration system	1	CentOS 7.3	2	1
Intranet-system of analysis of population data	1	CentOS 7.3	2	1
MongoDB request router	1	CentOS 7.3	1	2
MongoDB configuration server	1	CentOS 7.3	1	2
MongoDB shards	3	CentOS 7.3	2	2
Stage-server	1	CentOS 7.3	8	4
Jenkins Continuous Integration System	1	CentOS 7.3	4	2
DNS, DHCP	1	CentOS 7.3 or win 2012r2 if there is a subscription	4	2
Backup virtual machines	6	CentOS 7.3 or win 2012r2 if there is a subscription		

IV. THE CHOICE OF TECHNOLOGICAL SOLUTIONS FOR THE DEVELOPMENT OF THE CLIENT PART

Based on the information received, architecture synthesis was carried out to determine a set of loosely coupled components of the system, their connections, and the most effective ways of exchanging data.

To choose the programming languages and technologies solutions for the platform development, research and comparison were conducted. It was carried out in the context of the formed architecture, requirements and constraints.

For the programming languages and technologies analysis, reports and materials of services such as StackOverflow and GitHub, which are the most authoritative in the software development environment, were studied.

With regard to programming languages for browser applications, application delivery capabilities are evaluated without the need for additional software.

Frameworks are considered for their active application in projects, the size of the community of developers, the relevance of the task and the time on the market.

It should be noted that direct comparison of frameworks for development will not yield results, since each of them will allow to reach the final result. Nevertheless, a number of them should be considered more suitable due to better scalability, less training costs and more ready-made modules.

As a result of the consideration of JavaScript, Java applets and Adobe Flash platform from the point of view of applicability for code execution in the browser, it was found out that only JavaScript can be considered applicable.

This is due not only to the fact that JavaScript is used in many areas (client browsers, server part, mobile platforms), as well as the desktop applications. Java applet technology, like Adobe Flash technology, requires the installation of additional components in the user's system. Moreover, depending on the operating system and the browser, the installation and configuration process may vary.

Due to the great variability of the hardware and software in schools, the use of these two technologies is not advisable, since this can complicate the process of conducting mass research. It should be noted that in browsers on mobile devices Adobe Flash and Java applets are not supported. It's also worth noting that Adobe Flash rejects HTML5, which can work with multimedia (video and audio).

Thus, the choice in favor of the JavaScript language for the development of the client part becomes obvious and there is no alternative solution under the given conditions at the current moment. JavaScript is supported by all common browsers and is included in them by default.

Developing large Single-Page Applications (SPA) based on pure JavaScript on the client side is a difficult and inefficient process, so it is needed to use frameworks that define the application structure and have a basic set of components [23-25].

It should be noted that almost all modern frameworks have similar functionality and are able to solve the task. Thus, the choice should first of all be based not on the functional of the framework, but on the requirements and objectives within the framework of a particular project.

The most famous frameworks were selected for consideration, including Backbone.js, AngularJS 1, AngularJS 2, React, Ember.js, Vue.js and Polymer. Table II shows the advantages and disadvantages of these frameworks, taking into account the applicability to the developed platform.

Backbone.js is ill-suited for developing large projects, as there are no necessary components for implementing complex functionality. Thus, according to the authors of the article, the use of this framework is inexpedient in view of the fact that it

does not have sufficient functionality, and there are also alternative solutions.

TABLE II. ADVANTAGES AND DISADVANTAGES OF FRONT END FRAMEWORKS

Framework	Advantages	Disadvantages
Backbone.js [26]	Compact Simple structure Steep learning curve Rich documentation Supports REST	Does not support two-way data binding. Requires additional components to implement complex functionality. Bad for large projects
AngularJS 1 [27]	High popularity Steep learning curve Rich documentation Great community Many ready-made solutions It is part of the MEAN stack (MongoDB, ExpressJS, AngularJS, NodeJS) Supports REST High speed development Supports two-way data binding	It is believed to be outdated, since there is an AngularJS 2 Not compatible with AngularJS 2. The speed of work decreases with a sufficiently large amount of data
AngularJS 2 [28]	Rich documentation Great community Has a large number of functions Supports REST There are Angular Universal for solving problems of search engine optimization (rendering of pages on the server) Supports two-way data binding	Uses TypeScript to compile in JavaScript. Less steep learning curve compared to AngularJS 1. It is necessary to take many actions to solve even a small functional
React [29]	Compact High performance Good documentation Suitable for large and complex projects with a high degree of load	Requires additional implementation on the server for working with data (for example, Flux or Redux). Not supported by REST. Not compatible with libraries that modify the DOM. Less steep learning curve. Complex approach to development, unusual for beginners
Ember.js [30,31]	Rich documentation Large ecosystem Suitable for complex and large applications Supports REST Supports two-way data binding	It is considered to be monolithic in comparison with other frameworks. There is no reuse of components at the controller level. Less steep learning curve. Heavy structure. Too big for small projects
Vue.js [29]	Very rapidly growing popularity Steep learning curve Few dependencies Good performance Rich documentation Good ecosystem Supports two-way data binding	A fairly young framework. Developed mainly by one person. Not many projects were done. No "out of the box" REST support (there is an Axios library)
Polymer [32]	New and promising technology Web Components High speed of work	Too young solution. Great risks when using. Few ready solutions and examples. Less steep learning curve

Polymer is a library that is based on a fairly new technology Web Components technology. The W3C specification for this technology is not yet complete. There may be any problems with browser support, problems in stability of work, and also a large threshold of entry for developers. In this regard, the use of this framework was decided to be abandoned due to possible risks.

React, unlike others, is a library and does not allow to create a web application, since it is designed to create a View part and should work with data on the server, for example, in conjunction with Flux or Redux. Therefore, React is difficult to understand, has an unusual structure, which complicates the understanding of the application as a whole, and also has a large entry threshold for novice developers. According to the authors, React is more difficult to make a quick prototype and support the solution than on another framework.

AngularJS 1, AngularJS 2, Ember.js and Vue.js have two-way data binding, the ability to build large systems, good documentation and community. The choice will be made between these frameworks.

Ember.js has a complex project structure and a large entry threshold for novice developers, and in case of going beyond the standard use is cumbersome and not flexible. In addition, the framework is less popular than AngularJS and Vue.js.

Vue.js version 2 is currently the fastest growing popular framework, it took the best solutions from Ember.js, React and AngularJS, and also has good performance. Another important factor is that Vue.js does not support REST and requires an additional Axios library for this. In addition, the framework is young and is developed mainly by one person, so its use can lead to greater risks.

As a result, the most appropriate for developing a platform for psychological research is AngularJS 1 and AngularJS 2. AngularJS 1 is a fairly simple framework for mastering and understanding, has a low entry threshold with a rich set of functions. AngularJS 2 is a parallel project with AngularJS 1 and is developed separately. AngularJS 2 greatly complicated, for writing the simplest application requires much more action. In addition, it is written in TypeScript, which will require additional knowledge from the developers.

Taking into account what was written above, as well as the fact that the developed platform for psychological research has a limitation in resources and involves novice developers, and also the most appropriate solution for the current moment, according to the authors of the article, is AngularJS 1. In addition, AngularJS 1 has more popularity than other frameworks, according to GitHub and patent analysis [33].

V. SELECTION OF TECHNOLOGICAL SOLUTIONS FOR THE DEVELOPMENT OF THE SERVER PART

The development of the server part of the platform allows to choose from a fairly wide range of technologies, in comparison with the client part. This is primarily due to the fact that server technologies depend on the preferences of developers, equipment and requirements for the project, while client technologies are severely limited. The choice of technological solutions for the development of server

components is better to start not with programming languages, but with consideration of frameworks because they set the basic structure for the development of the application, as it was written above. Table III presents the features, advantages and disadvantages of the most suitable frameworks for the development of the server part of the platform.

TABLE III. ADVANTAGES AND DISADVANTAGES OF SERVER-SIDE FRAMEWORKS

Framework	Language used	Advantages	Disadvantages
Laravel, Symfony [34]	PHP	Steep learning curve A large number of PHP developers	Blocking IO calls PHP interpreter has low performance No paid support
Django [35]	Python	Steep learning curve Generating the administration panel for relational databases	Blocking IO calls Does not support NoSQL solutions out of the box
Ruby on Rails [36]	Ruby	Steep learning curve	In the development community, there are references to scaling problems under increasing load. Blocking IO calls
Express.js [33]	JavaScript (Node.js)	Not blocking by default (asynchronous) Steep learning curve	Long-term support of the project has difficulties (complexity of refactoring). Development in large groups can be difficult
Loopback [37]	JavaScript (Node.js)	Not blocking by default (asynchronous) Generating the Preview Panel and Working with the REST API Declarative approach to the generation of the REST API	The generated API does not contain methods for mass update of related entities
Play [38]	Scala / Java	Not blocking by default (asynchronous) Well scaled even with blocking code Strict typing simplifies refactoring	Slow compilation New versions of the framework require improvements in the final software
Vaadin [39]	Java	Contains the library of pre-made UI elements Frontend code is generated based on the server Strict typing simplifies refactoring	Blocking by default Slow compilation High threshold of occurrence The development of new UI elements is time-consuming There is no full control over the Frontend code
ASP .NET MVC [40]	C#	Strict typing simplifies refactoring	Lock-in to the Windows platform Need to purchase Windows Server licenses for deployment

Since it was determined that a high degree of project scalability is required, attention should be paid to non-blocking I/O frameworks. In this regard, it is necessary to exclude Laravel, Symfony, Django and Ruby on Rails from consideration. Also, due to the complexities of implementing non-blocking I/O and custom interfaces, the Vaadin framework is not suitable for the project.

ASP .NET MVC imposes additional restrictions on the infrastructure in the absence of significant advantages, so the framework should be excluded from further consideration. Thus, the main choice will be made between the Express.js, Loopback and Play frameworks.

An important factor is the programming language on which the framework is written. Express.js and Loopback are written in Node.js (JavaScript), while Play is developed on Java. In the case of JavaScript, a single syntax will be used for both the client and server parts. This will increase the effectiveness of the development of the platform, since the developer will need to know not two, but only one programming language, which is an advantage in the conditions of a small number of developers. In addition, it allows to combine part of the learning process and to reduce the overall threshold of entry, which will affect the time of training of new professionals who will participate in the development of the platform. Also, JavaScript is the most popular language in the world according to the statistics of such large services as GitHub and StackOverflow. In this regard, according to the article authors opinion, it is more expedient to use the Express.js and Loopback frameworks than Play Framework.

Of the remaining two frameworks, the choice in favor of Loopback is more appropriate for a number of reasons:

- Loopback offers a number of patterns that will help maintain the proper level of support for the code base as it increases;
- The framework is based on Express.js, which will enable all its functional components;
- Loopback offers functionality for simplified API generation, which greatly reduces the amount of labor involved in development.

All of the above mentioned, according to the authors of the article, is more significant than the steepest learning curve. Thus, the choice is stopped on the Loopback framework.

VI. DISCUSSION

With the use of previously described technologies the following key components of the platform were developed: API Server, Researcher Account, Test Player. Fig. 8 shows the page example of research project list. The example of test visualization for research participant is showed in Fig. 9.

The platform prototype is being tested in trial production. At the moment, the service functionality is partially implemented, but it is already possible to collect data using tests such as Dark Triad, Big Five and a number of techniques for assessing spatial thinking.

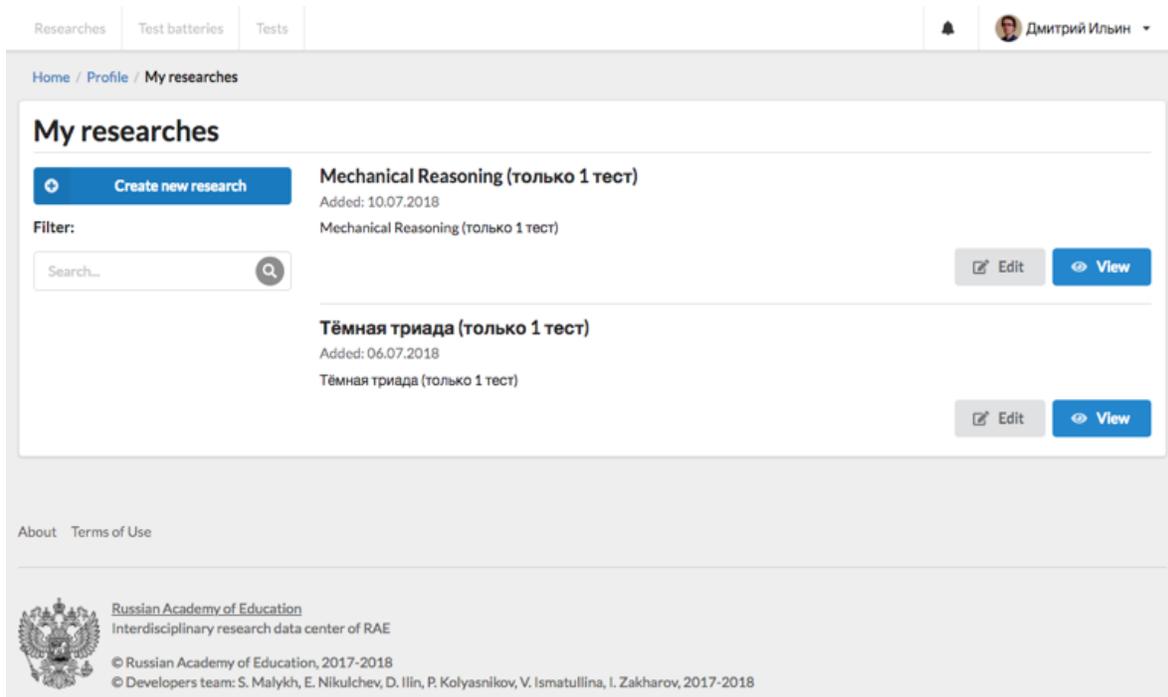


Fig. 8. Researcher Account Page.

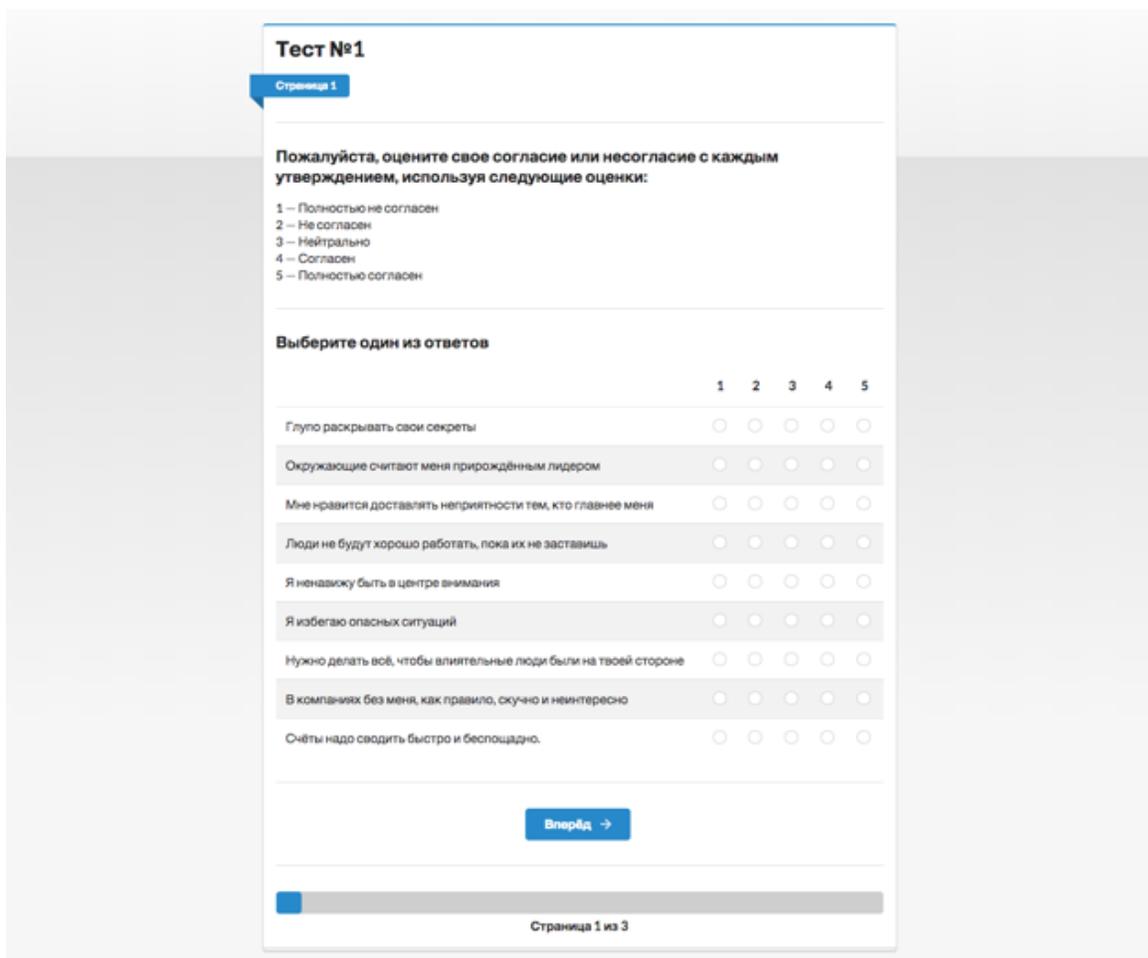


Fig. 9. Psychological Dark Triad Test Visualization.

VII. CONCLUSION

The guaranteed quality of disturbed system functioning, their effective and success work over a number of years, the scaling ability and connection to different platform are laid at the system designing stage. The chaotic development that was popular ten years ago belonged to the past. There are following reasons: often to the process of developing completion and software launching used technologies have become outdated. New approaches require careful and comprehensive documenting of the design and implementation process. The following concepts are widely used: ecosystem of programming languages, automated systems and development tools that allow designing software that is able to ensure the quality and reliability of the tasks.

The paper described a concrete example of the development process of the digital web-based platform for psychological research. The choosing of technologies for client and server part was showed. Currently, the system kernel is developed and being tested in one of the education institution for software support for psychological research. Developed system showed an adequate choice of software technologies. The system successfully operates in test mode.

In recent years, significant changes in the approaches to the study of problems of education have occurred in the world science. Rapidly developing high-tech methods of studying a person expand the possibilities for studying the mental development and learning of students. Research in this field becomes interdisciplinary and actively assimilates the approaches and methods of the whole series of other sciences. The use of new technologies of human study forms a huge amount of data even on small samples. Providing the conditions for training requires taking into account the age and individual psychological features of modern students. To solve this problem, population studies are needed, which are the basis for determining contemporary age norms of the student's mental development and national standardization of psychological diagnostic methods.

Significant individual differences exist in activities related to various forms of education (primarily, schooling). These differences are associated with both general cognitive abilities and private cognitive functions: motivation, emotions in response to the learning process, school and family environment. This determines the importance of researching individual differences for all fundamental learning problems. Each person has a unique genetic profile, which in turn forms individual psychological characteristics. The environment adapts to genes with the participation of parents, school and students themselves.

The complex nature of the interaction of genetic and environmental factors of individual differences at the psychological and psychophysiological levels requires the selection of adequate tools for conducting research.

The concept of the architecture of the web-based platform is formulated. A scalable multicomponent architecture is proposed. Scalability of the data warehouse is provided by technologies of sharding and replication. A list of virtual machines for deployment on the server has been generated.

Web-based platform is divided into server part (REST API server and data warehouse), public part (online test player, online test designer, researcher's personal profile and personal profile), private part (administration panel and data analysis panel) and external applications (desktop and mobile). In terms of security, the private part is used as a separate service on the intranet and has its own REST API. To implement the architectural aspects, the most appropriate technologies were chosen in the given task: JavaScript language that will be used to implement most of the software components, AngularJS framework for the client part of the online application, Loopback (Node.js) framework for implementing the API server that provides a single access point for all other components of the platform. The result of the techniques applying is implemented programmatically in the prototype platform, which is being tested in production trial.

ACKNOWLEDGMENT

The work was financed by the Ministry of Education and Science of Russia, project 25.13253.2018 / 12.1 "Development of the technological concept of the Data Center for Interdisciplinary Research in Education".

REFERENCES

- [1] H. Gilberstadt, R. Lushene, & B. Buegel. "Automated assessment of intelligence: The TAPAC test battery and computerized report writing," *Perceptual and motor skills*, vol. 43, no. 2, pp. 627-635. 1976
- [2] J.A. Naglieri, F. Drasgow, M. Schmit, & R. Velasquez. "Psychological testing on the Internet: new problems, old issues," *American Psychologist*, vol. 59, no. 3, p. 150-162. 2004
- [3] M. Buhrmester, T. Kwang, & S.D. Gosling. "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on psychological science*, vol. 6, no. 1, pp. 3-5, 2011.
- [4] D.J. Hauser, & N. Schwarz, "Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants," *Behavior research methods*, vol. 48, no. 1, pp. 400-4072, 2016.
- [5] F. Gino & S.S. Wilermuth, "Evil genius? How dishonesty can lead to greater creativity," *Psychological science*, vol. 25, no. 4, pp. 973-981, 2014.
- [6] C. Perilloux & R. Kurzban, "Do men overperceive women's sexual interest?" *Psychological Science*, vol. 26, no. 1, pp. 70-77, 2015.
- [7] D.N. Shapiro, J. Chandler & P.A. Mueller, "Using Mechanical Turk to study clinical populations," *Clinical Psychological Science*, vol. 1, no. 2, pp. 213-220, 2013.
- [8] R. B. Cattell, "Structured personality-learning theory: A wholistic multivariate research approach," Praeger Publishers, 1983.
- [9] V. Ismatullina, I. Voronin, A. Shelemetiya & S. Malykh, "Cross-cultural study of working memory in adolescents," *Procedia-Social and Behavioral Sciences*, vol. 146, pp. 353-357, 2014.
- [10] I. Zakharov, E. Nikulchev, D. Ilin, V. Ismatullina & A. Fenin, "Web-based Platform for Psychology Research," *ITM Web of Conferences*, vol. 10, p. 04006.
- [11] R.A. Stern & T. White, "NAB, Neuropsychological Assessment Battery: Attention Module Stimulus Book. Form 2," Lutz, FL: Psychological Assessment Resources, 2003.
- [12] I.M. Zakharov, I.A. Voronin, V.I. Ismatullina & S.B. Malykh, "The Relationship between Visual Recognition Memory and Intelligence," *Procedia-Social and Behavioral Sciences*, vol. 233, pp. 313-317, 2016.
- [13] I. Voronin, V. Ismatullina, I. Zaharov & S. Malykh, "The nature of the relationships between personality and cognitive characteristics," *SHS Web of Conferences*, vol. 29, p. 02043, 2016.
- [14] A. Chetverikov & P. Upravitelev, "Online versus offline: the Web as a medium for response time data collection. *Behavior research methods*," vol. 48, no. 3, pp. 1086-1099, 2016.

- [15] D. Venkatesan & S. Sridhar, "A novel programming framework for architecting next generation enterprise scale information systems," *Information Systems and e-Business Management*, vol. 15, no. 2, pp. 489-534, 2017.
- [16] Y. Chen, J. Huang, C. Lin & J. Hu, "A partial selection methodology for efficient QoS-aware service composition," *IEEE Transactions on Services Computing*, vol. 8, no. 3, pp. 384-397, 2015.
- [17] H. Zhang, M. Lu & T. Gu, "SOA software architecture extended modeling considering reliability information," In *2017 Second International Conference on Reliability Systems Engineering (ICRSE)*, IEEE, 2017, pp. 1-6. <https://doi.org/10.1109/ICRSE.2017.8030794>
- [18] S.B. Malykh & T.N. Tikhomirova, "Personality features and intellect: the nature of correlation." *Voprosy psichologii*, vol. 2, pp. 147-160, 2015.
- [19] S. Chattopadhyay, A. Banerjee & N. Banerjee, "A fast and scalable mechanism for Web service composition," *ACM Transactions on the Web*, vol. 11, no. 4, p. 26, 2017. <http://dx.doi.org/10.1145/3098884>
- [20] M. Barak & S. Ziv, "Wandering: A Web-based platform for the creation of location-based interactive learning objects," *Computers & Education*, vol. 62, pp. 159-170, 2013.
- [21] N. Venkateswaran & S. Changder, "Simplified data partitioning in a consistent hashing based sharding implementation." In *Region 10 Conference, TENCON 2017*, IEEE, 2017, pp. 895-900. <http://dx.doi.org/10.1109/TENCON.2017.8227985>
- [22] N. Venkateswaran & S. Changder, "Handling workload skew in a consistent hashing based partitioning implementation," In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017, pp. 1163-1169. <http://dx.doi.org/10.1109/ICACCI.2017.8125999>
- [23] E.K. Kristensen & A. Møller, "Type test scripts for TypeScript testing," In *Proceedings of the ACM on Programming Languages, (OOPSLA) 2017*, p. 90.1-90.25. <http://dx.doi.org/10.1145/3133914>
- [24] M. Dhok, M.K. Ramanathan & N. Sinha, "Type-aware concolic testing of JavaScript programs," In *Proceedings of the 38th International Conference on Software Engineering*, ACM, 2016, pp. 168-179. <http://dx.doi.org/10.1145/2884781.2884859>
- [25] S. Bae, J. Park & S. Ryu, "Partition-based coverage metrics and type-guided search in concolic testing for JavaScript applications," In *2017 IEEE/ACM 5th International FME Workshop on Formal Methods in Software Engineering (FormaliSE)*, IEEE, 2017, pp. 72-78. <http://dx.doi.org/10.1109/FormaliSE.2017.10>
- [26] A. Mardan, "Backbone.js and Parse.com," In *Full Stack JavaScript*. Apress, Berkeley, CA, 2015 pp. 121-136. https://doi.org/10.1007/978-1-4842-1751-1_5
- [27] W. Chansuwath & T. Senivongse, "A model-driven development of web applications using AngularJS framework," In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, IEEE, 2016, pp. 1-6. <https://doi.org/10.1109/ICIS.2016.7550838>
- [28] Y. Fain & A. Moiseev, "Angular 2 Development with TypeScript," Manning Publications Co., 2016. <https://dl.acm.org/citation.cfm?id=3133734>
- [29] D. Sheppard, "PWAs From the Start. In *Beginning Progressive Web App Development*," Apress, Berkeley, CA, 2017, pp. 209-240. https://doi.org/10.1007/978-1-4842-3090-9_11
- [30] A. Hamdy, O. Ibrahim & A. Hazem, "A Web Based Framework for Pre-release Testing of Mobile Applications," *MATEC Web of Conferences*, vol. 76, p. 04041, 2016.
- [31] J. Cravens & T.Q. Brady, "Building Web Apps with Ember.js: Write Ambitious Javascript," O'Reilly Media, Inc., 2014.
- [32] E.K. Kristensen & A. Møller, "Inference and evolution of typescript declaration files," *Lecture Notes in Computer Science*, vol. 10202, pp. 99-115, 2017. https://doi.org/10.1007/978-3-662-54494-5_6
- [33] E. Nikulchev, D. Ilin, G. Bubnov & E. Mateshuk, "Scalable service for predictive learning based on the professional social networking sites," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 5, p. 9-15, 2017.
- [34] S. Sinha, "How Request, Response Work in Laravel 5. In *Beginning Laravel*," Apress, Berkeley, CA, 2017, pp. 161-165. https://doi.org/10.1007/978-1-4842-2538-7_18
- [35] K. Lotfy & M.L. Hale, "Assessing pairing and data exchange mechanism security in the wearable Internet of Things," In *2016 IEEE International Conference on Mobile Services (MS)*, IEEE, 2016, pp. 25-32. <https://doi.org/10.1109/MobServ.2016.15>
- [36] Y. Zhang, G. Yin, Y. Yu & H. Wang, "Investigating social media in GitHub's pull-requests: a case study on Ruby on Rails," In *Proceedings of the 1st International Workshop on Crowd-based Software Development Methods and Technologies*. ACM, 2014, pp. 37-41. <https://doi.org/10.1145/2666539.2666572>
- [37] G. Rankovski & I. Chorbev, "Improving Scalability of Web Applications by Utilizing Asynchronous I/O," *Advances in Intelligent Systems and Computing*, vol. 665, pp. 211-218, 2016.
- [38] J. Hunt, "Play framework. In *A Beginner's Guide to Scala, Object Orientation and Functional Programming*," Springer, 2018, pp. 431-446. https://doi.org/10.1007/978-3-319-75771-1_38
- [39] K. Alexopoulos, S. Koukas, N. Boli & D. Mourtzis, "Resource Planning for the Installation of Industrial Product Service Systems," In *IFIP International Conference on Advances in Production Management Systems*. Springer, 2017, pp. 205-213. https://doi.org/10.1007/978-3-319-66926-7_24
- [40] A. Troelsen & P. Japikse, "Introducing ASP.NET MVC," In *Pro C# 7*. Apress, Berkeley, CA, 2017, pp. 1179-1221. https://doi.org/10.1007/978-1-4842-3018-3_29

Review of Prediction of Disease Trends using Big Data Analytics

Diellza Nagavci, Mentor Hamiti, Besnik Selimi
Faculty of Computer Science and Technologies
South East European University (SEEU)
Tetovo, Macedonia

Abstract—Big Data technologies promise to have a transformative impact in healthcare, public health, and medical research, among other application areas. Several intelligent machine learning techniques were designed and used to provide big data predictive analytics solutions for different illness. Nevertheless, there is no published research for prediction of allergy and respiratory system diseases. However, the impact of research and the finding of different cases is conducive to progress and further development of this. One of the goals of this paper is to devise a systematic mapping study, to explore and analyze existing research about disease prediction in healthcare information. According to the realized investigation of published research from 2012 up to today, we are focusing our research on studies that have been published around big data analytics. With this high number of secondary studies, it is important to conduct a review and provide an overview of the research situation and current developments in this area.

Keywords—Big data; algorithms; data analytics; healthcare; disease prediction; data mining

I. INTRODUCTION

Advancements in Information Technology have proven monumental in improving the quality of live throughout many dimensions. Medical Sciences is no exclusion to this relentless evolution. Internet, robots/ AI, and telemedicine have been very important in science when it is about adopting the medical science and profession with this trend; yet it is often argued that when it comes to critical thinking, no AI can beat the instincts of an experienced Doctor [1]. Insofar as ‘preparing for the future’ is concerned, Data Analytics have proven a highly reliable source of information in a plethora of sciences, and since ‘all data is equal’ for the AI, prognosis of health conditions and eventual epidemics using Big Data is particularly attainable, and immensely important.

Based on this premise, we are focusing our research on finding a suitable data mining algorithm for using Big Data to predict diseases. We believe that analyzing big data sets will lead to finding causes that lead to respiratory disease and allergies to populations in the future.

In this paper, we first introduce the methodology used, and the research questions defined. Secondly, we give a classification scheme of the fields of interest, big data, data mining, data sets, and optimization. Afterwards, we provide answers to five research questions and two are proposed as future aim research. On the discussion part, the time series of papers relating to health information field of interest have

been included. As a future research we have proposed big data analytics and data mining learning algorithms.

II. METHODOLOGY

The main goal of this mapping study is to define each step of research fields answer the research questions based on analyzed articles [2]. The idea was to collect a series of publications in the field of interest, to determine the coverage of the research field. For categorization of reports and different results, we needed the structure and for that we had taken a systematic approach. It shows results by using the visual summary and a map. We have used different research questions that must be defined to obtain these objectives in a systematic manner. The main purpose of a structured mapping study is to present an overview of a certain research area as well as to identify research gaps. A systematic literature review is another kind of secondary study that answers specific research questions by identifying, analyzing and interpreting relevant evidence. The process begins with the definition of research questions, from which we can arrive to a research scope. The next step is to conduct the actual search by retrieving all papers that may be remotely related to the field. Then comes the screening of the papers, with the objective to filter all the relevant papers. The classification scheme is based on key wording by reading the abstracts. In the end is to extract the data and to show results.

A. Research Questions and Search Strategy

- What is the principle of interest discussed in the papers?
- What type of framework is used for Big Data?
- How publications have evolved over time? What the research and publication trends are?
- Which methods are used previously?
- Which algorithms are used for processing Big Data?

Most of the explored research publications are extracted from digital libraries as IEEE-Xplore, ACM, and some of articles are from Springer and IJRCCE. The search strings in Table I are used to search in digital libraries.

Large number of articles appeared on different search strings. Have been selected just the ones that we saw reasonable to include as more appropriate and help achieve our goal. Most of papers have been published in recent years. In the Fig. 1 are shown the number of published papers by

year. The papers that have been published last year are from the first half of 2017. From the selected papers, further analysis is conducted and in this study papers related to Map-Reduce and Big Data, health information, data mining and prediction are included. As a result, after removing duplicates and irrelevant papers, only 119 articles have been filtered.

TABLE I. NUMBER OF PAPERS BY MAIN FIELD OF INTEREST

No.	Search String	No. of papers
SS1	((("Abstract": Map Reduce) OR "Abstract": big data) AND "Abstract": health information)	217
SS2	((("Map Reduce") OR " big data ") AND health information) AND data mining	133
SS3	((("map reduce framework ") OR " big data ") AND health information) AND data mining) AND prediction	69

III. CLASSIFICATION SCHEME

The classification scheme is presented in three columns where we include the main fields of interests related to the research (Fig. 1). The Big Data analytics are the main fields on which we will focus to propose a more suitable algorithm in the future. The field of interests are defined in the first column to come up with framework types. And the third column is for different big data processing algorithms. Based on the analysis from the collected papers, we have found the research gap in the ‘big data analytics’, in which further contribution is expected from the research community. The third column shows different algorithms for processing big data is random walks that needs to be fulfilled in order the future goals be verified. More details are presented in research questions in the results part. Additionally, we classified papers according to the field of interest.

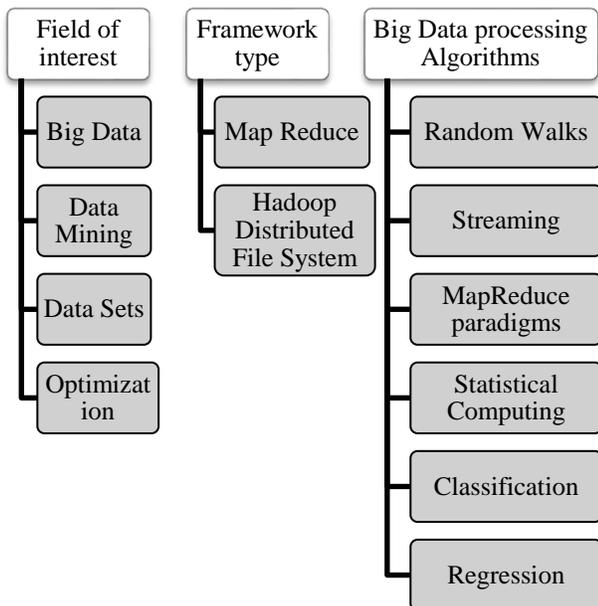


Fig. 1. Classification Schema.

IV. RESULTS

Each article was classified into the categories of each facet in order to answer the six research questions. The results of the systematic mapping study are presented as follows.

A. RQ1: What are the Principle of Interest Discussed in the Papers?

This question deals with the main field of interest that is investigated in each of the papers. We are interested in Big Data, but since we used several search strings, we got several results. In order to answer this question, we created the ‘Field of Interest’ classification for the papers.

From Table II, we can observe that about 58% of the papers have as main focus in the big data that refers to data sets and flows large enough that pose significant challenges when using commonly available tools and infrastructures for collecting, managing and processing the data within a tolerable amount of time.

The second most mentioned area is data mining with almost 21%. This category of papers includes tasks actions like as data extraction techniques and driving manual tools needed to adapt to new technologies to overcome time constraints.

TABLE II. NUMBER OF PAPERS BY MAIN FIELD OF INTEREST

Field of interest	Number of papers	Percentage
Big Data	69	58%
Data mining	25	21%
Data sets	15	13%
Optimization	10	8%

B. RQ2: What Type of Framework is used for Big Data?

A lot of proposals that we have searched during the research have been focused in MapReduce approach [3]. MapReduce paradigm has been used to implement Classification techniques. The data that is processed and disseminated in a cloud computing infrastructure is very convenient and very effective to accelerate the process of knowledge generation.

Here are two types of frameworks:

1) Hadoop–provides its own file system called HDFS (Hadoop Distributed File System). To find the solution of the data text in a Hadoop shows us that all the data are performing parallel operation in different clusters.

Hadoop also will keep the multiple copies of data in case of hardware failure [4].

2) The second framework is A MapReduce that consists of two functions: map and Reduce. These two functions take a set of important pairs/value data and generate a set of output key/value pairs when a Map Reduce job is given to the cluster. The job is divided in two pieces into map tasks and reduce tasks, where each Map task will process one block of input data. A Hadoop cluster takes slave nodes to execute Map and

reduce task. The slave node it is capable to except number of map and Reduce tasks and execute simultaneously. A slave node sends a signal to master node in a given period of time. To accept the signal it will request the master node to slave. The Map function waits for worker node that shows input key and value pairs outside of the block [5].

From the investigation of our papers, we found out the results in Table III, according to which the mass of the papers use Map Reduce framework (45%). The rest of the papers deal with Hadoop, about 55%.

TABLE III. NUMBER OF PAPERS BY FRAMEWORK TYPE

Framework Type	Number of papers	Percentage
Map Reduce	54	45%
Hadoop	65	55%

C. RQ3. How Publications have Evolved Over Time? What the Research and Publication Trends are?

While studying the year of publication for each paper, we notice that the time ranges between 2013 and 2017. The majority of the papers (31.93%) have been published in 2017. In fact if we look at graph in Fig. 2, we notice that the lot of papers increases from year to year.

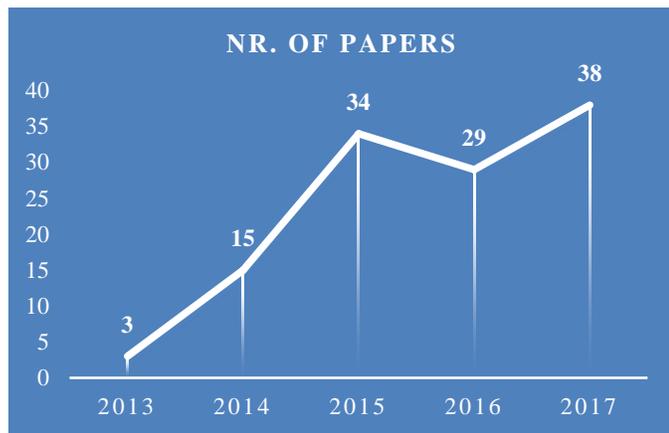


Fig. 2. Number of Papers per Year.

D. RQ4: Which Methods are used Previously to Process a Big Data Analysis?

In recent years, big data analytics is so important to health care. Big data analytics is a very broad area that deals with the collection, storage and analysis of immense data sets to trace the unknown patterns and other key information. The data that are very important it can help us recognize the data that are integral component to the future business decisions. The second research question is concerned with the different methods that have been used in big data analytics.

Omar El-Gayar and PremTimsina [6] have proposed a model on Evidence Based Medicine and Big Data Analytics. The main purpose of the system is to improve the cost across the applications of business intelligence big data analytics.

Samir El-Masri et al. [7] in this paper the authors have designed a model for clinical Decision Support System. The model is described as an Adaptive Evidence based Medicine. Using this model, the patient data from Electronic Health Record (EHR) was collected in a data warehouse. The Clinical Practice Guidelines (CPG's) were taken and CPG rules were generated using an automated converter. These rules were applied on the data obtained from the warehouse and the standardized data was stored in the knowledge base. The inference engine processed the questions from the physician and searched the knowledge base for the most applicable guideline. The model lacked in handling distributed nature of the process, that is, the CDSS that were geographically distributed could not interact with each other. The rules required to be standardized before processing, there increasing the time complexity of the system.

Sankaranarayanan. S and Pramananda Perumal. T. [8] have invented a model for Diabetic prognosis using Data Mining techniques. The two major data mining algorithms Apriori and FPGrowth were applied on Diabetes Mellitus dataset to generate association rules. Frequent item sets were mined after which rules were generated using support and confidence threshold values. This model was not generic to diagnose variety of diseases. Accuracy of the prediction was not guaranteed, and the model was not scalable to support voluminous health records.

Mohamed Abouzahra et al. [9] have implemented a model on integrating data from Electronic Health Record (HER) to improve Clinical Decision making for Inflammatory Bowel disease. This model accumulated fragmented data of a patient that can take from different EHR systems. Analytical techniques were applied to this data to identify useful patterns. These techniques were based on physicians' input, literature, and existing guidelines to identify possible relationships between different components of patients' data. Predictive methods were used to predict future outcome of the patient and facilitated the diagnosis of disease. The patients' privacy and information security issues were not treated properly.

E. RQ5: Which Algorithms are used for Processing Big Data?

One important part of our research is to find out which algorithms are used for processing big data analytics. J. Qiu et al. has presented different machine learning algorithms for big data processing [10]. The first one is representation learning or feature learning which deals with learning data representations that make the data analysis process easier. It is found that the performances of the machine learning algorithms are strongly influenced by the selection of data representation (or features) [11]. Feature selection (variable selection) techniques are used to find those features of data which are most relevant for use in model construction. Feature extraction techniques transform the high dimensional data into a low dimensional space. In space metric learning, the function of distance is constructed to calculate the distance between different points of a data set. Table IV represents a list of some of the algorithms that are used in different research papers. The authors mentioned about another hot learning technique called deep learning in their paper.

TABLE IV. DIFFERENT ALGORITHMS SUMMARY

Algorithms	Used
Random Walks Distributed Hash Tables, Bulk Synchronous Parallel (BSP)	Random walk is designed to address wide range of problems in mobile and sensor networks. For every machine to know that information resides is used the hash table. The BSP computer is compiled of a set of processors connected by a communication network
CART, Recursive Partition Trees	Decision tree algorithms
K- Nearest neighbor, Bayesian, SVM, ANN, K-means,	A survey was done on the various machine Algorithms for classification, prediction and modelling
MapReduce, Linear regression	The main objective was to improve the accuracy of rainfall forecasting.

V. DISCUSSION

After analyzing the 119 papers, our focus is to explore and to design an algorithm that will analyze and make prediction from the data sets, in different platforms. The idea is that in modern big data research, the suitability of different algorithms is solely dependent on the data characteristics. Therefore, there is a need for further in-depth analysis to find the suitable supervised and unsupervised machine learning algorithms to derive meaningful facts and actionable insights from HIS data.

J.L. Berral-Garcia has presented a paper describing the frequently used machine learning algorithms for big data analytics [12]. Several algorithms are used for performing modeling, prediction and clustering tasks. Decision tree algorithms (like CART, Recursive Partition Trees or M5), K-Nearest neighbors algorithms, Bayesian algorithms (using Bayes theorem), and Support vector machines (SVM), Artificial Neural Network, K-means, DBSCAN algorithms, etc are presented in this paper. Several execution frameworks - Map-Reduce Frameworks (Apache Hadoop and Spark) were also mentioned. The implementations of the previously discussed algorithms are made available to the public through different tools, platforms and libraries such as R-cran, Python Sci-Kit, Weka, MOA, Elastic Search, Kibana etc. M. U. Bokhari et al. presented a three layered architecture model for storing and analyzing big data [13]. The three layers are data gathering layer, data storing layer and data analysis & report generation layer. In order to gather and handle the huge volume of big data coming from high speed sources such as sensors or social media, a cluster of high speed nodes or servers are kept in the data gathering layer. The data storage layer is responsible for storing the big data. The Hadoop Distributed File System (HDFS) can be used for data storage [14]. Principal Component Analysis, Singular Value Decomposition and tensor-based approaches are useful for feature extraction. For feature selection, filter-based and wrapper-based methods are helpful. All these are dimensionality reduction techniques. The authors compared different techniques for performing data mining tasks. Logistic

regression, cox regression, local regression techniques are simple to interpret, but are prone to outliers. The authors discussed about the useful platforms for big data analytics. Apache Hadoop, IBM Platform, Apache Spark Streaming, Tableau, and other visual analytics tools are highly impactful platforms for providing big data analytics solutions. Two real world case studies such as integrative /omics data for the improved understanding of cancer mechanisms, and the incorporation of genomic knowledge into the EHR system for improved patient diagnosis and care were done to discuss the usefulness of biomedical big data analytics for precision medicine. Multi-omic TCGA [15] data and EHR data were used to conduct this study. Since we wanted to find the gap of where and how we can design an algorithm that will analyze and predict the data sets, in different platform. We will consider it for future goals and analysis.

VI. FUTURE RESEARCH

Our main objective in the recent future will be to analyze the different approaches and concepts previously used, determining which algorithms could potentially be appropriate for finding causes of respiratory diseases and allergies. These algorithms will be applied to data health information about patients and check what kind of prediction can be derived, how accurate are the predictions of each of these algorithms. Based on this, we want to derive a method that could provide satisfactory results in term of predicting disease trends.

Although we base on previous research, it should be remarked that it is a first attempt to suggest a concrete and detailed algorithm that will be implemented in the Kosovo Health Information System.

Two research questions proposed as future research are:

- Where does health big data come from?
- What value will give this algorithm in HIS?
- How can healthcare systems benefit from big data analytics?

VII. CONCLUSION

Big Data presents a unique discipline, which should have the most important role in the latest technology developments. We presented the different algorithms and technologies that are used in predicting diseases.

Although the main objective of this research is to apply existing algorithms in the prediction of a different disease, it is important to support it by practical example, limited to the data that can be made available on Kosovo Health. The collection, filtering, normalizing and processing of the data itself is an important problem – hence our focus on data mining and big data processing techniques.

In addition, it will serve as a recommended model for research and for further development of this field of research.

REFERENCES

- [1] Athmaja S., H. M. (2017). A SURVEY OF MACHINE LEARNING ALGORITHMS FOR BIG DATA ANALYTICS. International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).

- [2] Hirak Kashyap, H. A. (2014). Big Data Analytics in Bioinformatics: A Machine Learning Perspective. IEEE.
- [3] Isaac Triguero, D. P. (2014). A MapReduce solution for prototype reduction in big data classification. IEEE.
- [4] Krishna, P. R. (2015). Big Data Search and Mining. Springer India . IEEE.
- [5] Matthew Herland, T. M. (2014). A review of data mining using big data in health informatics. IEEE.
- [6] Megan Sheeran, R. S. (2017). A framework for big data technology in health and healthcare. . IEEE.
- [7] Mohamed Abouzahra, K. S. (2012). "Integrating Data from EHRs to Enhance Clinical Decision Making: The Inflammatory Bowel Disease Case". Proceedings of 27th International Symposium on Computer-Based Medical System.
- [8] Nadiya Straton¹, R. R. (2017). Big Social Data Analytics for Public Health: Comparative Methods Study and Performance Indicators of Health Care Content on Facebook. IEEE International Conference on Big Data (BIGDATA), 2772-2777.
- [9] Omar El-Gayar, P. D. (2014). "Opportunities for Business Intelligence and Big Data Analytics In Evidence Based Medicine". IEEE Trans, pp .749-757. .
- [10] Pramananda Perumal, T. S. a. (2014). "Diabetic prognosis using Data Mining methods and techniques". Proceedings of ICICA, Coimbatore, India.
- [11] Prof. Sharmishta Desai, S. R. (2016). VERY FAST DECISION TREE (VFDT) ALGORITHM ON HADOOP. IEEE.
- [12] Samir El-Masri, S.-S. (2012). "An Adaptive Evidence Based Medicine System Based on a Clinical Decision Support System". Science Series Data Report.
- [13] Sayali D. Kadam, P. D. (2016). Big Data Analytics- Recommendation System with Hadoop Framework . IEEE.
- [14] Shim, K. (2012). MapReduce Algorithms for Big Data Analysis. Proceedings of the VLDB Endowment, , Vol. 5(12),.
- [15] Y. Bengio, A. C. (2013). "Representation Learning: A Review and New Perspectives. IEEE.

The Role of Hyperspectral Imaging: A Literature Review

Muhammad Mateen, Junhao Wen, Nasrullah, Muhammad Azeem Akbar
School of Big Data and Software Engineering,
Chongqing University,
Chongqing, 401331, China

Abstract—Optical analysis techniques are used recently to detect and identify the objects from a large scale of images. Hyperspectral imaging technique is also one of them. Vision of human eye is based on three basic color (red, green and blue) bands, but spectral imaging divides the vision into many more bands. Hyperspectral remote sensors achieve imagery data in the form of hundreds of adjoining spectral bands. In this paper, our purpose is to illustrate the fundamental concept, hyperspectral remote sensing, remotely sensed information, methods for hyperspectral imaging and applications based on hyperspectral imaging. Moreover, in the forensic context, the novel methods involving deep neural networks are elaborated in this paper. The proposed idea can be useful for further research in the field of hyperspectral imaging using deep learning.

Keywords—Deep learning; electromagnetic spectrum; hyperspectral imaging; imaging spectroscopy; multispectral imaging; remote sensing

I. INTRODUCTION

In hyperspectral, the term “hyper” means “too many” and it refers to the huge amount of measured wavelength bands. Hyperspectral images are used to provide sufficient spectral information to recognize and differentiate spectrally distinctive materials. Optical analysis techniques are used to detect and identify the objects from a large scale of images. Hyperspectral imaging technique is one of them. Vision of human eye is based on three basic colors (red, green and blue) bands, but spectral imaging divides the vision into many more bands. Hyperspectral remote sensors achieve imagery data in the form of hundreds of adjoining spectral bands.

The collective data is used to obtain a constant spectrum for each imagery pixel mentioned in Fig. 1. After tuning the sensor, terrain and atmospheric belongings are applied [1]. These imagery spectrums can be analyzed with laboratory or field reflectance spectra to know and map surface materials such as specific kind of plants or indicative minerals with ore deposits.

Imaging spectrometers are instruments used to produce hyperspectral images. The remote imaging and spectroscopy are two basic technologies used to develop the hyperspectral sensors. Spectroscopy is a field of study about light emission or reflection from different materials and the changes occurred in energy with wavelength.

In the field of optical remote sensing, spectroscopy acts

with the spectra of sunlight that is scattered by objects at or inside the earth. Remote images are planned to capture and calculate the light returned from adjoining areas on the surface of earth. Hyperspectral imaging can be applied to various applications including medicines, biogeochemistry, biophysics, industrial monitoring and remote sensing to collect the information for analysis.

Chinese Academy of Science played a vital role in the field of hyperspectral imaging and developed two outstanding imagers, one of them is known as Push broom Hyperspectral Imager (PHI) and another one is Operative Modular Imaging Spectrometer (OMIS). In 2000, another invention about imaging came out in the form of Hyperspectral Digital Camera (HSDC) which supports limited spectral bands with high quality of spectral resolution [2]. HSDC plays a flexible role for different observation objects and applications including environmental and agricultural monitoring.

Analysis of Hyperspectral imager can be performed by two different ways one of them is perspective of spectral analysis and another is based on image processing. It is more important that the data should be well organized before selection of any kind of approach. In spectroscopic analysis, the spectra should be extracted by region of interest that is usually calculated by three different ways such as threshold an image with single waveband, ratio or difference image. In image processing, few images are selected from the collection of images for rapid computation. Selection of those images is based on the importance of their wavelength for shifting carefully. In Fig. 2 [3] for careful shifting, spectra provides an option for peaks and valleys whether in [4] based on original or preprocessed format.

There are several other methods to achieve the same goal, for example, partial least square regression and principal component analysis. Additionally, some data compression techniques such as singular value decomposition and Fourier transform are used for the process of more images to increase the ability of hyperspectral imaging [5]. After the achievement of healthy data, the next step is to make dependent standardized models. Before implementation of Chemometric algorithm, it is compulsory to overcome the noisy data to increase the quality of signals. Moreover, image processing includes filtering and binning, which can increase the quality of data. In calibration model, the same routine is followed by spectroscopy and spectral analysis as shown in Fig. 2.

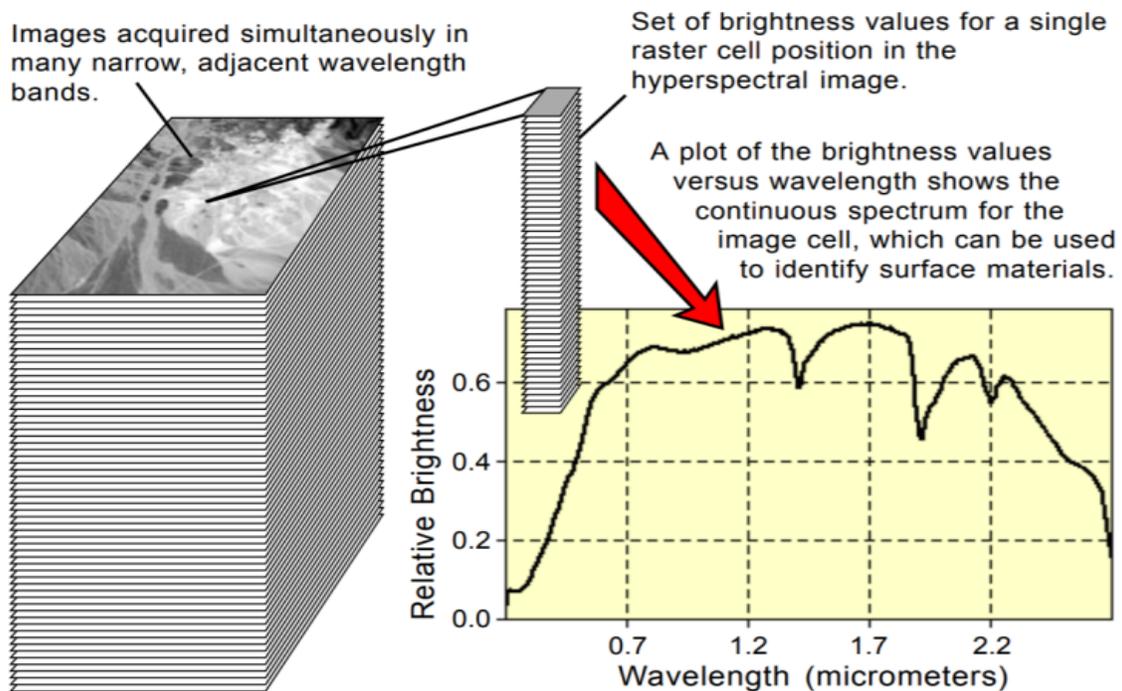


Fig. 1. Measurement of hyperspectral remote sensor.

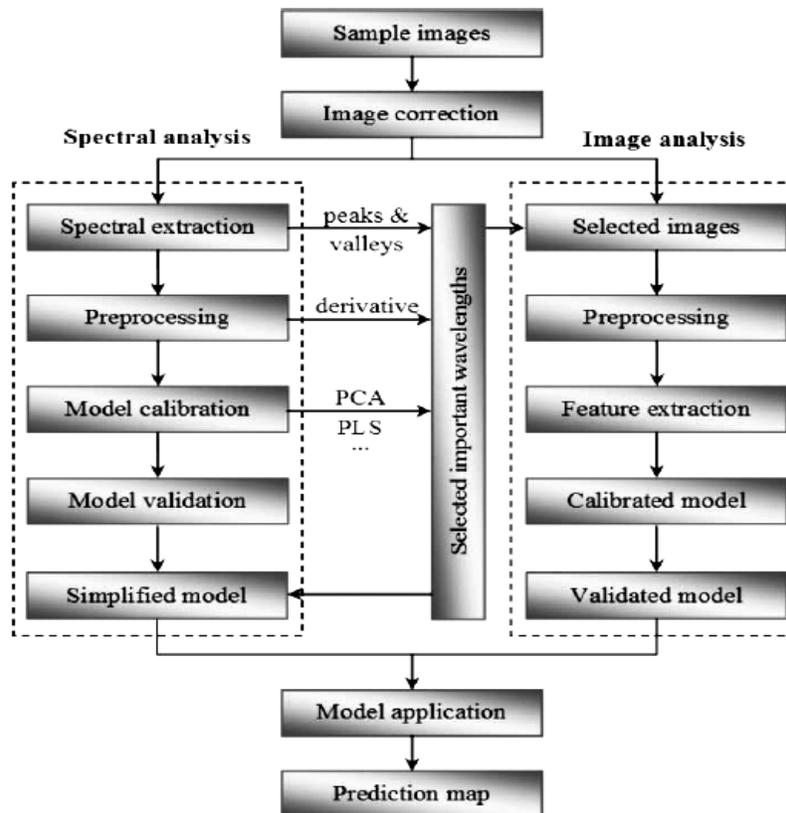


Fig. 2. Flowchart of hyperspectral imaging processing.

In this model different techniques such as principal component analysis, partial least square discriminate analysis, fisher's linear discriminate technique, support vector machine,

artificial neural network and fuzzy inference are offered [3]. As a result prediction map can be progressed with the use of post-processing techniques and the information is achieved.

The rest of the paper is organized as: related work of the study is discussed in Section II. Study findings are reported in Section III, and Section IV covers the future work of the study. The applications of hyperspectral imaging are explained in Section V. Finally, the findings are concluded in Section VI.

II. RELATED WORK

Light emitted by sun strikes on the molecules, causing in absorption and reflection of light which depends on the structure or architecture of molecules. The wavelength of absorption or reflection of atomic bonds and molecule makes detection or identification of a specific object. To collect the amount of light reflected by a specific object or earth surface is used for identification of that object. There are different types of sensors used to collect the scattered data including truck-mounted sensor, airborne sensor and satellite-based sensor. With the help of fast computers and development of sensors, the immense research [6] is being carried out in the field of "Hyperspectral remote sensing".

In [7], Sun, L et al. introduced a novel approach based on supervised classification for hyperspectral imaging. Supervised classification model includes a spatially weighted random fields and spectral data reliability. To progress the quality of classification, the real labels of training data samples are fixed as a constraint in the model of proposed approach. In [8], Sun, L. et al. introduced noise (deadlines, strip noises, impulse and Gaussian) reduction technique for hyperspectral imaging with the use of low-rank representation on the spectral image. According to the proposed technique spectral space of hyperspectral image recline in the low-rank subspace. The low rank representation based on spectral space was oppressed by nuclear norm of image with the spectral dimension. It causes potential for the removal of deadlines, strips, sparse noise and Gaussian noise on the level of each band.

In the field of image classification, high dimensionality of imaging data is a challenge. Hyperspectral imaging data also contains high dimensionality, so the spatial-spectral classification methods are demands for computational view point. In [9], Wu, Z. et al. introduced a novel technique for parallel implementation of spatial-spectral classification on the basis of adaptive random fields. In this technique, logistic regression classifier is used for spectral information. In this research, graphics processing units are used for parallel implementation. GPU sets the work allocation, input and output between the CPU and the graphics processing unit having fully computational control of GPUs with low latency and high bandwidth of shared memory.

In [10], Sun, L. et al. introduced a novel approach for noise reduction in the hyperspectral images. Noise reduction technique is based on the super pixel-based low level rank of representation for hyperspectral imaging. Under the model of a linear mixture, the hyperspectral cube is considered as a low rank in the spectral area, which is divided HSI data into the sub-matrices of lower ranks.

In 1985, a technique named as "Imaging spectrometry" was used for the remote sensing of the earth. In addition, there

was a motivation for the development of digital image processing methodologies by multispectral data analysis [11]. In the early stage, Goetz et al. [6] defined Hyperspectral remote sensing as "The collection of images in dozens to hundreds of adjacent registered spectral bands so that for every cell of an image a glowing spectrum can be calculated." According to this definition, spectral regions of VIS (visible), NIR (Near Infrared), SWIR (Shortwave Infrared), MWIR (Midwave Infrared), LWIR (Longwave Infrared) and UV (Ultraviolet) are covered.

A. Hyperspectral Remote Sensing

The hyperspectral sensor is used to study soil science, geology, mining, land use, and hydrology to map and identify geometric and chemical patterns of land. Information obtained by hyperspectral sensor is used to identify valuable minerals and petroleum. Imaging spectroscopy is a technique used to absorb specific features of chemical bonds in minerals. Material mapping includes water, ice, snow, mineral mixtures, vegetation, environmental materials, atmospheric gases, and man-made materials. Hyperspectral airborne scanners were used in 1998 to identify rocks and soils containing hydrocarbon [12]. This experiment was named as "Pro Smart Experiment" managed by German Aerospace Center to test the Hyperspectral "HyMap" designed by the Australian Company Integrated Spectronics Ltd. In Hyperspectral data analysis, several factors are involved to make it complex for refined techniques and algorithms.

To demonstrate the hyperspectral data analysis, themes are managed in six main areas including un-mixing, data fusion, target detection, physical parameter retrieval, classification and fast computing based on image processing, machine learning, and signal processing. According to [13], researchers described all the areas one by one in the paper. In the data fusion, various strategies are defined including restoration of the signal to noise ratio (SNR) and spatial resolution, spectral data fusion to overcome high spectral redundancy, spatial data fusion used to enhance the resolution of imaging sensor at the sub-pixel level. There is another sensor named as "Dual push broom hyperspectral sensor" which contains two various devices (covers different wavelengths) that are normally attached to the similar bench. "Co-registration of pixels" and "Peer of the field of view" are problems lead by this configuration called Bore sight Effect. To observe and detect spectral phenomena, A. Brook et al. [14] proposed a prior to its correction with the use of Bore sight Effect that is used to provide supplementary information not located in the corrected images.

B. Remotely Sensed Data

Hyperspectral sensors collect the data from the surface of the earth with the help of different parameters. It is important to measure the accurate values of identified objects. There are a lot of methods and techniques, developed to measure the values of vegetation. Some techniques show that the results can be obtained in highly accurate form but few techniques show the results in moderately accurate form.

In [15], Mutanga et al. studied on a small number of chemicals, for example, chlorophyll and another kind of pigments. The pigments including carotenes and chlorophyll

are identifiable with 80 per cent accuracy in sample fields. Nitrogen can also be detected with accuracy. But the other nutrients namely magnesium, calcium, sodium, potassium, and phosphorus have less notified by Mutanga et al. in his research. Even though this is very expensive for projects to analyze with a lot of samples, the impending of remote sensing is not totally oppressed. In the [16], research related with Geoinformation science and earth analysis for modeling and management of environment designed a remote sensing data model in which “spot vegetation” concerned with geo-referenced and “Corine land cover data” concerned with study area that provides qualitative and quantitative information of earth surface. This research work used datasets with some characteristics including geometric accuracy better than 100 m, thematic accuracy greater than 85 percent, and spatial resolution of 30m.

For the development of South African Imaging Spectroscopy Landscape, an overview is explained in brief about analysis techniques and challenges [17]. According to African research and application, indicators can be derived by the full width of the spectral response for example slopes, integrals, and curve derivatives. The purpose is to identify data source from the imaging spectroscopy, in which spectral attributes are required to notify an application of interest. According to [18], object-based classification approach is used to classify the remotely sensed data. In 2000, Walter and Fritsch introduced a concept using multispectral remote sensing data to automatically revise the geographical information system databases. In object-based classification, not only single pixel is classified but the group of pixels is also classified based on geographical information system database. This approach depends on two main steps, the first step is based on supervised likelihood classification and the second step is based on matching of existing GIS objects with classified objects to detect the change occurred or to identify the faulty data. This approach develops and makes the better result of the classification. At the end, the result is obtained in the form of fully classified, partly classified and not found by the use of a threshold, iteratively defined by a user. In [19], a surface bidirectional reflectance model has been designed. The model is applicable for heterogeneous surfaces and follows a semi-empirical approach. There are three parameters used to describe the nature of the surface and these parameters are used in the effective algorithm of correction and processing the remote sensing data. The model based on surface bidirectional reflectance is collection of two basic processes, one of them is to diffuse reflection component which explains the geometrical structure on surface and also understands the shadow effects while the second process defines the volume scattering properties obtained by isolated surfaces. Invisible and near-infrared spectral bands, model and situ annotations demonstrate the better association for common investigated surface types. The model is good to reduce the largely unwanted variations related to surface bidirectional results in remotely sensed data.

U.C. Benz et al. [20] proposed object-oriented analysis associated with fuzzy methods to implement information and explained workflow from remote sensing imagery to graphical information system. During proposed method, software named

as “eCognition” is used to offer the suitable connection between remote sensing imagery and graphical information system. Thus object-oriented technique can provide powerful automatic and semiautomatic evaluation to allocate practiced knowledge to GIS.

III. DISCUSSIONS

Hyperspectral imagery process follows some methods to accomplish the proposed task. These methods contain study area, which means where experiment is performed; field sampling, meaning the sample data gathered from study area; Chemical analysis, which means evaluation of chemical compounds and Spectral processing, which means to perform operation on sample of spectra to overcome noisy data, statistical formulas are used for prediction of parameters. D. Liao et al. [21] proposed method to visualize hyperspectral image in normal color by the coordination of HSI and high resolution image through multiple alignments. Manifold alignment identifies the matching color points and displays it into pair wise alignment, while the spaces between them play the role of bridges. Association of hyperspectral imaging and RGB image generates a spatial image in the natural format. This approach has advantage of flexibility that’s why it can be applied for different scenario. The same approach is also well known because of user’s customization, in which user can scan visual resultant bands according to the specified interest. Chemometric and spectroscopy areas provided methods and analysis tools including “Partial Least Squares Regression Analysis” and “Principal Component”, which are useful for the process of hyperspectral image [22]. Workflow of hyperspectral image processing is different from the workflow of color image processing but both data types are multivariate and multidimensional. Hyperspectral imaging methods such as image acquisition, spectral and spatial preprocessing, dimensionality reduction, calibration, feature extraction and selection are used for HSI processing as shown in Fig. 2. Push broom line scanner is commonly used as distributed hyperspectral imaging sensor. Push broom scans all wavelength data points with the same spatial coordinates. In the hyperspectral imaging system, calibration is a method that ensures the repeatability and accuracy of results gained by hyperspectral imaging data. Calibration is a procedure that connects wavelengths with band numbers. Spatial calibration is a method of measuring the correlation of each image pixel such as meters. Spatial preprocessing is also a method used in hyperspectral imaging to minimize the noisy data from the images. Spatial post-processing is considered more valuable than spatial preprocessing because prediction and classification images are common images that need spatial manipulation, interpretation and pattern identification. Spatial sampling and Region of Interest (ROI) is a basic method of hyperspectral imaging used to mention the study area or location of interest for hyperspectral sensors. Spatial sampling can also be considered as background masking produced by binary images [22]. Transformation to absorbance is also used during reflectance or transmittance of data for analysis [23]. Hyperspectral imaging contains noisy data during scanning a large scale of image. To reduce noisy data, de-noising algorithm is performed in the spectral domain of hyperspectral imaging [24]. Feature extraction is also essential method used

to transform the existing features to a set of new features. Selection of subset from input feature without transformation is called selection extraction. It is also called feature subset selection. Mosaic based images [25] which are merged into a particular hyperspectral image, is used for the data analysis and classification algorithms.

In [26], affinity propagation method is used for the selection of bands in hyperspectral imaging. In this method centered based clustering approach is used to classify the similar color bands. AP is applied by factor graph and then operation performed by the centralized data points through message passing until an appropriate set of bands is achieved. The presentation of band selection is analyzed by the classification of a pixel inside the specific image. Classification of pixels is better performed by affinity propagation instead of all other approaches. Unlike conventional methods of clustering named as K-means, agglomerative clustering, a proposed technique (AP) obtained better results of band selection through message passing method. Hyperspectral imaging contains massive information of colored bands. Collection of colored bands can have noisy data that is considered to remove before processing. During preprocessing manual band removal technique is used to remove the noisy data. In [27], problem related with MBR is identified, because during MBR an important data can also be removed. For this problem, a proposed technique is introduced to automatically select the noisy bands instead of MBR selection. In the proposed technique, first wavelet reduction is applied to de-noise the bands of specified image and then Affinity Propagation approach is used to classify the

representative bands from the noisy data with a smart way. To overcome noisy data in an efficient way, two sensors are applied, so the experimental results show that proposed technique is better performed than Manual Band Removal technique.

To overcome the noisy data from hyperspectral images is also introduced by [28], a proposed technique of sparse representation based on noise reduction method. This technique depends on a non-noisy component which can be sparsely decayed over redundant dictionary rather than a noisy component. The paper shows the correlation of spectral-spatial structure of HSI by use of three dimensional blocks instead of two dimensional patches for sparse representation. Gaussian and Poisson noise models are collectively used for signal-dependent and signal independent noises in hyperspectral imagery. The proposed technique is good for virtual and real data sets of HS remote sensing.

Naganathan et al. [29] proposed meat tenderness for the satisfaction of consumers. Purpose of this proposal was to build up and test a near infrared/visible hyperspectral imaging system to guess tenderness of cooked beef by hyperspectral images. For this purpose, a push-broom hyperspectral imaging system associated with diffuse floodlighting system was designed and standardized. Three tenderness categories including tender, intermediate and tough were used to identify the features of meat tenderness. Statistical textual features obtained from Slice Shear Force (SSF) analysis are used in “canonical discriminate model” for prediction. The results signify that hyperspectral imaging played a vital role in the prediction of meat tenderness.

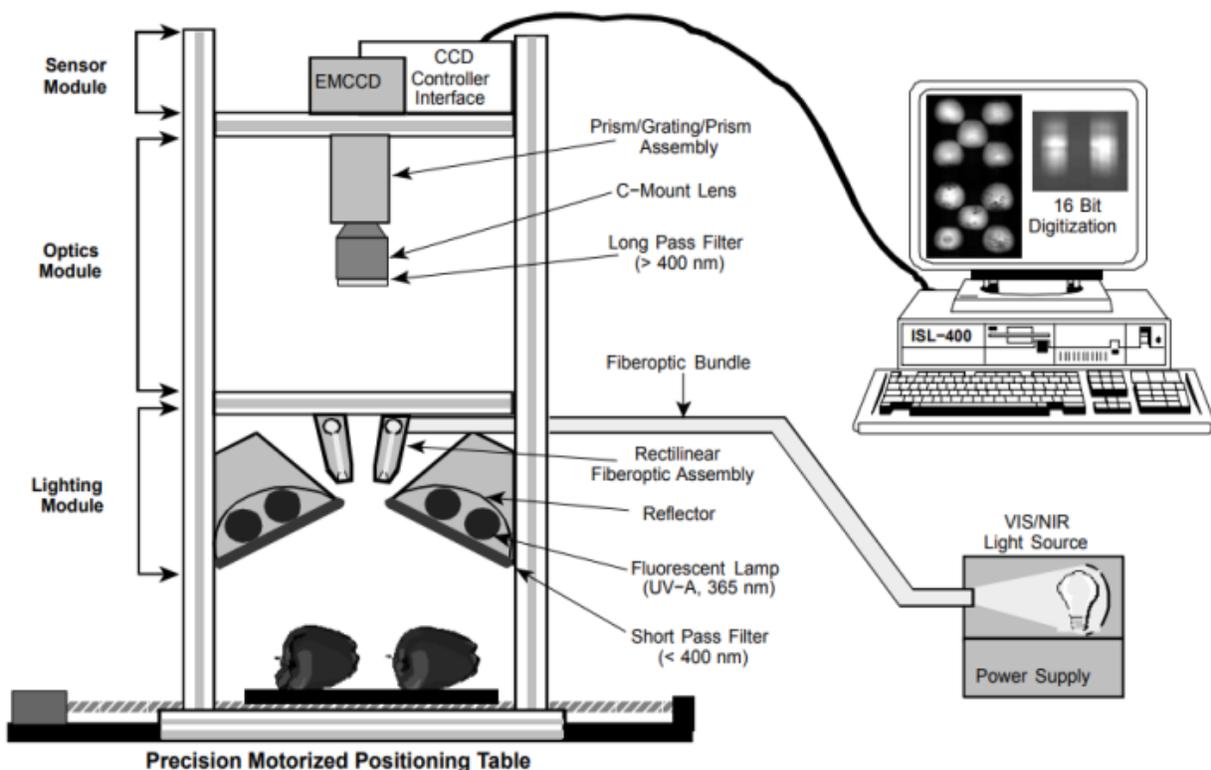


Fig. 3. Schematic diagram of the ISL hyperspectral imaging system.

TABLE I. DETECTION RESULTS OF SKIN TUMOR

Image	No. of Real Tumors	No. of Detections	No. of False Positives	No. of Missing
1	3	3	1	0
2	3	3	0	0
3	1	1	1	0
4	3	2	1	1
5	4	4	1	0
6	2	1	1	1
7(2)	3	3	1	0
8(2)	7	5	3	2
9(2)	3	3	3	0
10	12	6	0	6
Total	41	31(76%)	12(28%)	10(24%)

Hyperspectral Imaging system is used for detection of skin tumor [30], in which first of all the carcass of chicken is analyzed for the usage of spectral information as shown in Fig. 3, and results are applied for the detection of skin tumor as shown in Table I.

There was a little deficiency about detection that tumor less than 3 millimeters in diameter could not be identified. This step in research opens a gate for the detection of tumors by hyperspectral imaging and can also be improved for computational complexity.

In the field of poultry farms, hyperspectral imaging is used to design a rapid, accurate and non-distractive method to detect the embryo and fertility improvement in eggs [31]. A near infrared spectral imaging system was introduced for the identification of fertility and early embryo improvement. For the experiment, totally 174 white shell eggs were used including 18 infertile eggs and 156 fertile eggs incubated for 4 days. During the inspection by hyperspectral imaging, eggs were categorized into two parts one for fertile and other for infertile and the dataset of each category is different with the day of incubation. Gabor filter used to extract the image texture information of eggs. K means clustering technique is also applied to cluster out the data and top results were achieved as 84.1% at day 4, 81.8% at day 3, 74.1% at day 2 and 100 % at day 0. The result shows that last three bands were used for detection because of maximum response of spectral transmission and can be applied for real-time detection system for early embryo and fertility of chicken eggs. In the field of hyperspectral imaging, a large number of researchers are focused to design different kinds of spectral-spatial techniques while remaining are paying attention towards experimental development [32]. To evaluate the hyperspectral imaging classification techniques, it needs the reliable plan of testing including suitable standardized data sets, sampling procedure for training and testing data, and fair

analysis standard [33]. J. Liang et al. [34], proposed random sampling approach for spatial-spectral techniques to decrease the overlapping between testing and training samples and offer more intention and precise evaluation. Random sampling approach [35] is often used because it cares all tagged data evenly and every sample is picked with the similar likelihood.

In [36], there is a review of hyperspectral imaging analysis with different techniques. In [36], Artificial Neural Network, Auto-Encoder, Stack Auto-Encoder, Conventional Neural Network and deep learning is discussed for the analysis of Hyperspectral Imaging. After comparison of all these techniques and getting results, the author mentioned that deep learning outperformed in the analysis of hyperspectral imaging among all other stated techniques.

Furthermore in [37], there is a new model named as R-VCANet designed in the context of deep learning for classification of hyperspectral imaging. R-VCANet model is a combination of Rolling Guidance Filter and Vertex Component Analysis Network. R-VCANet is useful when there is limited sampling for feature extraction of hyperspectral data. R-VCANet is based on natural characteristics of HIS data, spectral properties, and spatial information. Hence the method proposed in [37] has performed better for hyperspectral image classification, especially when the sampling labels are limited.

Geophysics plays a vital role to study about graves detection after some changes are occurred in Buried graves. To detect the clandestine graves, local environment and different types of soils can be observed and data can be collected which is used for analysis. So it is important to become familiar with the equipment and hardware that is used for data acquisition and data analysis. John J. Schultz [38] used ground-penetrating radar for detection of clandestine graves. To use the GPR, it is also important to adjust the antenna for the best frequency of forensic work, which depends on spodosol environment. Electromagnetic induction meter is also used to locate the clandestine graves, but EMI is limited for forensic research, so the lack of published research with the use of EMI meter to detect clandestinely buried bodies became deem in the forensic context. As mentioned above that EMI is limited for forensic research, a new idea is introduced to do better in the forensic field. According to that, buried remains can be easily detected by hyperspectral imaging. By hyperspectral image, a large scale of an image can be achieved and with the application of deep neural networks, it can be classified in a more précised way.

For un-mixing, classification and target detection of hyperspectral images, there are frequently used methods includes, sparse linear models, Gaussian mixture models, latent linear models, ensemble learning, directed graphical models, support vector machines, linear regression, logistic regression, Gaussian models, clustering and deep learning. The summary of all these methods is highlighted in Table II with referred research articles.

TABLE II. METHODS FOR HYPERSPECTRAL IMAGING

Methods	Un-mixing		Classification		Target Detection	
	Linear	Non-linear	Spatial-spectral	Pixel-wise	Target	Anomaly
Sparse Linear Models	spatial-spectral [39]		feature extraction [40]	feature extraction [41]	spatial-spectral [39]	
Gaussian Mixture Models			[42]	un-supervised [43]		[44]
Latent Linear Models	[45]		feature extraction [46]	dimensionality reduction [47]		
Ensemble Learning			[48]	transfer learning [49]		[50]
Directed Graphical Models	sub-pixel mapping [51]	spatial-spectral [52]				
Support Vector Machines	end member extraction and sub-pixel mapping [53]	[54]	[55]	band selection [56]	[57]	[58]
Linear Regression	[59]	[60]				
Logistic Regression			semi-supervised [61]	band selection [62]		
Gaussian Models				transfer learning [63]	[64]	[65]
Clustering				un-supervised [66]		[67]
Deep Learning			un-supervised feature learning [68] supervised feature learning [69]	supervised feature learning [70] un-supervised feature learning [71]		[72]

IV. FUTURE WORK

To locate clandestine graves is a challenge for government (forensic department) after any victims. There is a lot of traditional ways to locate and detect the clandestine graves, for example with the help of forensic trained dogs, ground penetrating radar, electromagnetic induction meter, and hyper spectral sensor, in the forensic context. There are lots of researchers who are doing work to facilitate the detection of clandestine graves. In case of sudden disaster, the government forensic agencies play a vital role to detect the victims. Forensic archaeologist and anthropologists face a lot of challenges in forensic context.

One problem which Saudi emergency responders have to deal with is that sometimes heavy rains result in the flood. Victims of floods get buried deep in sand at ranges from 50cm up to 2 meters. The soil under which the victims are buried is likely clayey loam and poorly drained, and it can be covered with water for some time (few days). The emergency responders focus their search operations in areas where victims are likely to be located such as valleys (e.g., Alhayer valley) shown in Fig. 4.



Fig. 4. Heavy Rain Flood in Alhayer Valley (Saudi Arabia).

However, the suspected areas are typically very large rendering conventional ways of grave detection such as using forensic detection dogs, ground-penetrating radars, and thermal imaging inefficient, let alone the huge cost and large human resources required in the search operations.

In this case, the idea of examining the emission and biogeochemistry of gases from graves and their detection through remote sensing, in particular, hyper-spectral images sounds like an interesting idea. A hyperspectral imager mounted on an Unmanned Airborne Vehicle (UAV) can fly over the suspected area and acquire imagery which may contain the absorption features (signature) of the gases of interest predominantly CH₄, N₂O, and CO₂. The higher pore air concentrations in graves and emission of methane, carbon dioxide and nitrous oxide to the atmosphere imply the existence of graves in that specific location, which can be associated with the missing victim. In contrast with dry sandy soil which is aerobic and therefore conducive to methane consumption and/or oxidation, the search area is clayey loam and poorly drained soil, and thus susceptible to methane production. This increases the chance of detecting gases of interest using hyper-spectral imaging techniques.

For better understanding about hyperspectral imaging using deep learning, we demonstrated the process of feature extraction using the deep neural network. In Fig. 5, there is a data patch of hyperspectral image for input layer and on the next number of hidden layers; hyperspectral image is further divided for feature extraction process. After the process of hybrid feature extraction results can be achieved from the output layer. The outcome can be in the form of 1 or 0 that shows the presence and absence of buried remains respectively. According to the current research trends, deep learning has become an advanced and robust technique to extract the features of hyperspectral image as compared to the traditional feature extraction techniques.

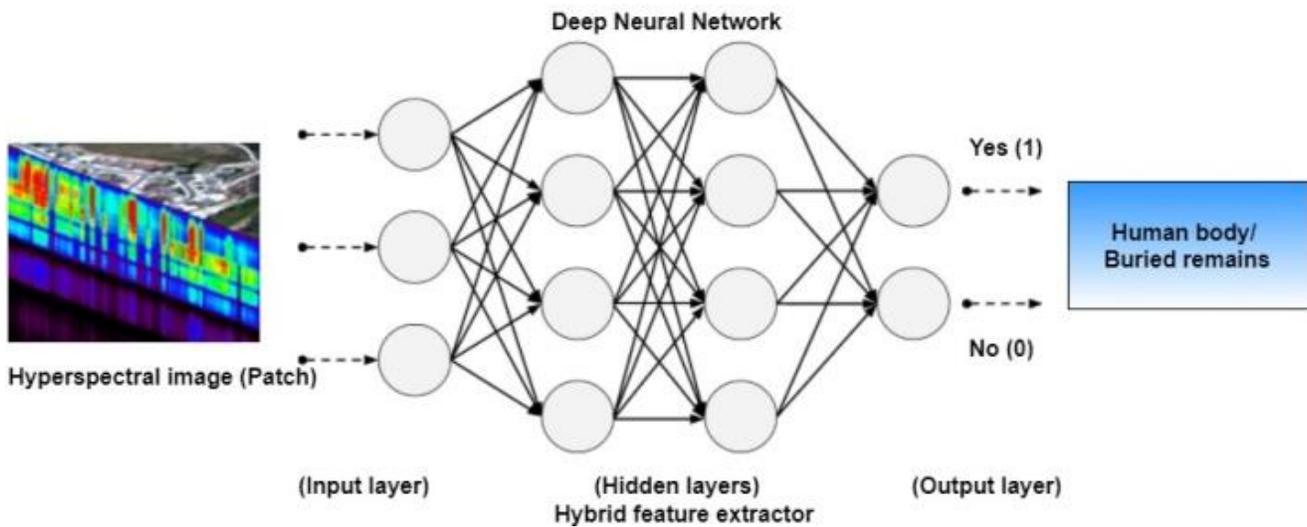


Fig. 5. Hyperspectral imaging using deep neural network.

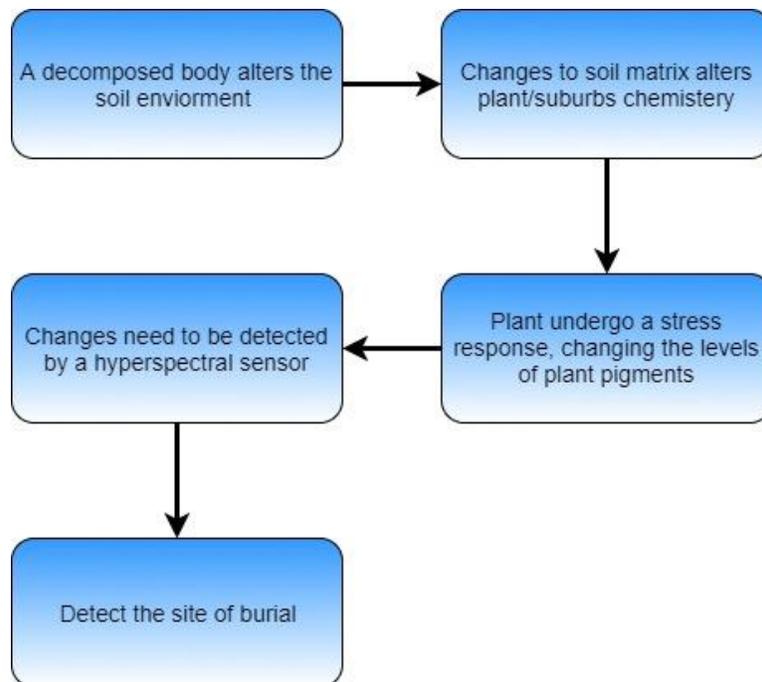


Fig. 6. Flowchart of future study.

The detection of clandestine and previously unknown burial sites is of interest to governments to start rescue operations. The detection of clandestine graves is an emerging tool in hyper-spectral remote sensing. In literature, studies have shown that it is possible to use hyperspectral remote sensing techniques in detection of mass graves. For this purpose, a UAV with hyperspectral sensors covering the visible to the shortwave infrared range was used to collect the imagery. The flow chart of the proposed technique is shown in Fig. 6.

Hardware Required for Completion of Project:

- Hyperspectral Sensor (data acquisition)
- An Unmanned Airborne Vehicle (UAV)

- High spec PC/Laptop for experiment and data analysis

Techniques/tools used for Completion of the Project:

- Matlab 2017
- Hybrid Feature Extraction
- Deep Neural Network (Classification)

V. IMPLICATIONS OF HYPERSPECTRAL IMAGING

Hyperspectral imaging is widely used in various applications such as Agriculture, Mineralogy, Surveillance, Physics, Astronomy, Chemical Imaging, and Environment. In the agricultural industry, diseases cause a serious loss for the economy. For suitable agriculture, it is important to monitor

plants and trees for diseases. To detect diseases on early stage can save the plants and trees from further loss, Hyperspectral sensor can assist to control the virus by organized methods such as fungicide applications, disease-specific chemical applications and pesticide applications [73]. The hyperspectral imagery is used in land cover mapping. Land cover classification associated with the nature of land such as grassland, forest, concrete pavement, and sand etc. On the other hand, Land use indicates the human use of land, for example, industrial, residential and agricultural zone. Land use classification is an application used to classify the land surface such as Geometric correction, ground truth, and maximum likelihood. Land cover change detection is another application used to detect the changes in the earth surface. It can be identified with the comparison of existing image and the updated one. Changes on earth have two main types. One is seasonal change and other is annual change, in seasonal change forests or plants get changes according to the season, but in annual changes, new things take place for example developments on the earth or deforested place. Global vegetation map is another application of hyperspectral imaging. According to this application global vegetation index data contains information of normalized difference vegetation index (NDVI). NDVI is updated on the weekly basis, so it can also contain some noisy data. Water is an essential part of life. Water quality monitoring is an important application of hyperspectral imager. Quality of water can be analyzed due to its greenish or yellowish shaded color. To measure the surface temperature of sea hyperspectral imager can also

perform the services. The hyperspectral sensor can also provide thermal information for a short interval of time over a large amount of area shown in Fig. 7. Brightness Temperature is a sensor used to detect the temperature of objects.

Every object has different emissivity which discharges electromagnetic energy. The value of emissivity is nearly equal to 1 and remains constant as compared with the temperature of the earth. Snow covered area is also the detectable area for hyperspectral sensors. During snow survey, snow water equivalent has been planned with the approximation of snow-covered area. Height measurement can also be calculated by Hyperspectral imager with the matchup of stereo images. In [74], there are two techniques to measure the height of objects, one is already defined as stereo matching and other is based on analytical plotters.

Fruits are the important part of food because it provides vitamins and energy to human's body. To gain healthy and fresh fruits is also a challenge. Hyperspectral imager provides the applications to monitor the quality of fruits. Peng and Lu [75] developed a reflective system to identify apple firmness and solids contents with the use of steady object stage. With the help of optical fiber and focusing lenses, 2-dimensional hyperspectral images were obtained. Huang and Lu [76] introduced another quality attribute analyzed by hyperspectral imaging is known as mealiness. Haung and Lu analyzed the association between apple mealiness and reflective hyperspectral line images. The mealiness of apple was calculated by the solidity and ripeness.

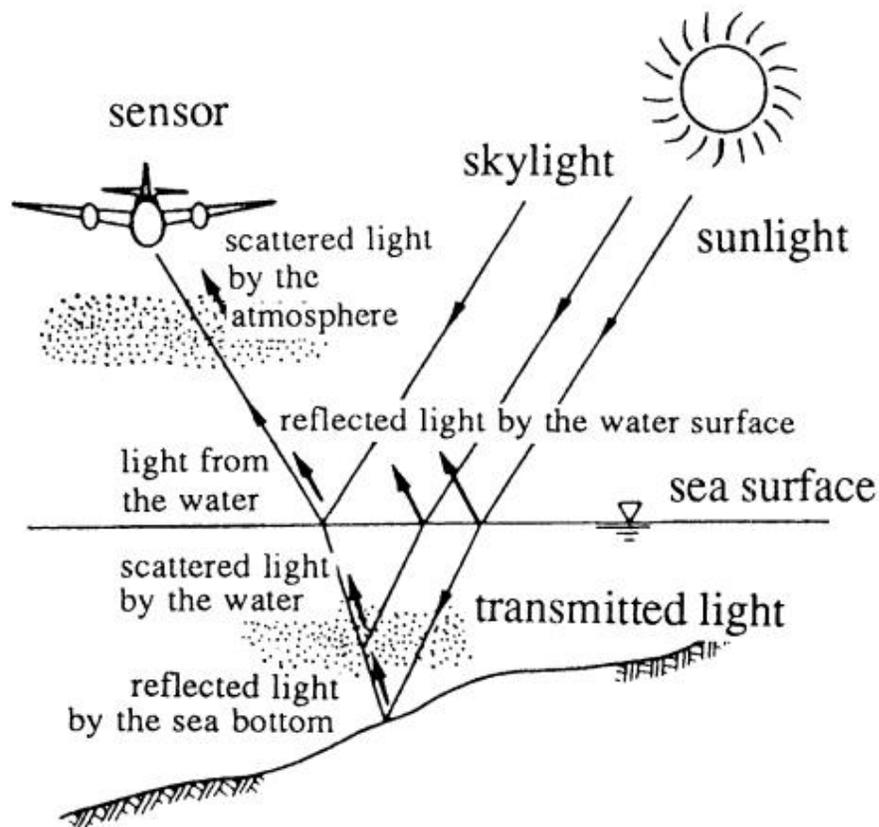


Fig. 7. Incident light into a sensor on the sea.

Remote sensing of vegetation is widely used in the world to check the quality of vegetables. The important application of hyperspectral imaging about vegetables including mushrooms, onions, cherry tomato, and spinach leaves. Ariana and Lu [77] designed a VIS-NIR hyperspectral imaging system joining reflective mode and transmitted mode collectively, while with the use of moving transport proposal. The system was able to identify internal issues of pickles and cucumbers which were impossible to visualize by the human eye. Hyperspectral sensors can also be used to monitor the quality of meat such as pork, chicken, and beef. In the context of pattern recognition, marbling is used to detect the quality of meat. To extract the marbling, a proposed technique used a wide line detector implemented by pattern recognition [78]. Results are achieved with the help of standard marbling charts and classified with the accuracy of 99 per cent. Hyperspectral imaging is also applied for seafood such as prawn, cod and smoked salmon. Near Infrared (NIR) Hyperspectral System was designed to detect the grains diseases on the early stage with the use of mathematical modeling framework. This system was based on supervised classification [79]. In [80], Khan et al. studied on advanced developments in the field of HIS Analysis. Details of this review based on fundamentals of Hyperspectral Imaging including spatial resolution, spectral resolution, temporal resolution, feature extraction. Secondly, authors explained the advance applications of hyperspectral imaging such as food quality, safety, security, remote sensing, especially in the field of forensic documents examination to study about intrinsic and extrinsic elements. After a long discussion about hyperspectral imaging, we have come to know that hyperspectral imaging is an active area for research, through which new ideas can be implemented and the world can be changed.

VI. CONCLUSION

In this paper, we described the fundamental concept, hyperspectral remote sensing, remotely sensed information, methods for hyperspectral imaging and applications based on hyperspectral imaging. We introduced a better approach to detect the buried remains with the use of deep neural networks for feature extraction. Finally, we focused on the use of hyperspectral imaging in different applications. However, hyperspectral imaging technology using deep learning is gradually becoming a great opportunity for researchers, in the field of biomedical, vegetation, especially in the forensic context.

ACKNOWLEDGMENT

This research was supported by the Basic and Advanced Research Projects in Chongqing, China under Grant No.61672117.

REFERENCES

- [1] Ahmadi, S. B. B., Nanehkaran, Y. A. and Layazali, S. Review on hyper-spectral imaging system. *Int. J. Sci. Eng. Res.*, 4, 5 2013), 253-258.
- [2] Tong, Q., Zhang, B. and Zheng, L. S. Hyperspectral remote sensing technology and applications in China(2004).
- [3] Feng, Y.-Z. and Sun, D.-W. Application of hyperspectral imaging in food safety inspection and control: a review. *Critical reviews in food science and nutrition*, 52, 11 2012), 1039-1058.
- [4] Wang, H., Peng, J., Xie, C., Bao, Y. and He, Y. Fruit quality evaluation using spectroscopy technology: a review. *Sensors*, 15, 5 2015), 11889-11927.
- [5] Bonnier, F., Bertrand, D., Rubin, S., Venteo, L., Pluot, M., Baehrel, B., Manfait, M. and Sockalingum, G. Detection of pathological aortic tissues by infrared multispectral imaging and chemometrics. *Analyst*, 133, 6 2008), 784-790.
- [6] Lanaras, C., Baltasavias, E. and Schindler, K. ADVANCES IN HYPERSPECTRAL AND MULTISPECTRAL IMAGE FUSION AND SPECTRAL UNMIXING. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 40(2015).
- [7] Sun, L., Wu, Z., Liu, J., Xiao, L. and Wei, Z. Supervised spectral-spatial hyperspectral image classification with weighted Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 53, 3 2015), 1490-1503.
- [8] Sun, L., Jeon, B., Zheng, Y. and Wu, Z. Hyperspectral image restoration using low-rank representation on spectral difference image. *IEEE Geoscience and Remote Sensing Letters*, 14, 7 2017), 1151-1155.
- [9] Wu, Z., Shi, L., Li, J., Wang, Q., Sun, L., Wei, Z., Plaza, J. and Plaza, A. GPU parallel implementation of spatially adaptive hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11, 4 2018), 1131-1143.
- [10] Sun, L., Jeon, B., Soomro, B. N., Zheng, Y., Wu, Z. and Xiao, L. Fast Superpixel Based Subspace Low Rank Learning Method for Hyperspectral Denoising. *IEEE Access*, 6(2018), 12031-12043.
- [11] Goetz, A. F., Vane, G., Solomon, J. E. and Rock, B. N. Imaging spectrometry for earth remote sensing. *Science*, 228, 4704 1985), 1147-1153.
- [12] Moser, G., Serpico, S. B. and Benediktsson, J. A. Land-cover mapping by Markov modeling of spatial-contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE*, 101, 3 2013), 631-651.
- [13] Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N. and Chaussonot, J. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and remote sensing magazine*, 1, 2 2013), 6-36.
- [14] Brook, A. and Ben-Dor, E. Advantages of the boresight effect in hyperspectral data analysis. *Remote Sensing*, 3, 3 2011), 484-502.
- [15] Mutanga, O. and Skidmore, A. K. Continuum-removed absorption features estimate tropical savanna grass quality in situ. *City*, 2003.
- [16] Diaz-Delgado, R. and Pons, X. Spatial patterns of forest fires in Catalonia (NE of Spain) along the period 1975-1995: analysis of vegetation recovery after fire. *Forest ecology and management*, 147, 1 2001), 67-74.
- [17] Mutanga, O., Van Aardt, J. and Kumar, L. Imaging spectroscopy (hyperspectral remote sensing) in southern Africa: an overview. *South African Journal of Science*, 105, 5-6 2009), 193-198.
- [18] Walter, V. Object-based classification of remote sensing data for change detection. *ISPRS Journal of photogrammetry and remote sensing*, 58, 3-4 2004), 225-238.
- [19] Roujean, J. L., Leroy, M. and Deschamps, P. Y. A bidirectional reflectance model of the Earth's surface for the correction of remote sensing data. *Journal of Geophysical Research: Atmospheres*, 97, D18 1992), 20455-20468.
- [20] Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I. and Heynen, M. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of photogrammetry and remote sensing*, 58, 3-4 2004), 239-258.
- [21] Liao, D., Qian, Y., Zhou, J. and Tang, Y. Y. A manifold alignment approach for hyperspectral image visualization with natural color. *IEEE Transactions on Geoscience and Remote Sensing*, 54, 6 2016), 3151-3162.
- [22] Park, B., Yoon, S.-C., Windham, W. R., Lawrence, K. C., Kim, M. S. and Chao, K. Line-scan hyperspectral imaging for real-time in-line poultry fecal detection. *Sensing and instrumentation for food quality and safety*, 5, 1 2011), 25-32.

- [23] Burns, D. A. and Ciurczak, E. W. Handbook of near-infrared analysis. CRC press, 2007.
- [24] Cen, H. and He, Y. Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science & Technology, 18, 2 (2007), 72-83.
- [25] Yoon, J. H., Sheremata, S., Rokem, A. and Silver, M. A. Windows to the soul: vision science as a tool for studying biological mechanisms of information processing deficits in schizophrenia. Frontiers in psychology, 4(2013), 681.
- [26] Qian, Y., Yao, F. and Jia, S. Band selection for hyperspectral imagery using affinity propagation. IET Computer Vision, 3, 4 (2009), 213-222.
- [27] Jia, S., Ji, Z., Qian, Y. and Shen, L. Unsupervised band selection for hyperspectral imagery classification without manual band removal. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5, 2 (2012), 531-543.
- [28] Qian, Y. and Ye, M. Hyperspectral imagery restoration using nonlocal spectral-spatial structured sparse representation with noise estimation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6, 2 (2013), 499-515.
- [29] Naganathan, G. K., Grimes, L. M., Subbiah, J., Calkins, C. R., Samal, A. and Meyer, G. E. Visible/near-infrared hyperspectral imaging for beef tenderness prediction. Computers and electronics in agriculture, 64, 2 (2008), 225-233.
- [30] Kim, I., Kim, M., Chen, Y. and Kong, S. Detection of skin tumors on chicken carcasses using hyperspectral fluorescence imaging. Transactions of the ASAE, 47, 5 (2004), 1785.
- [31] Liu, L. and Ngadi, M. Detecting fertility and early embryo development of chicken eggs using near-infrared hyperspectral imaging. Food and Bioprocess Technology, 6, 9 (2013), 2503-2513.
- [32] Lu, D. and Weng, Q. A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing, 28, 5 (2007), 823-870.
- [33] Friedl, M., Woodcock, C., Gopal, S., Muchoney, D., Strahler, A. and Barker-Schaaf, C. A note on procedures used for accuracy assessment in land cover maps derived from AVHRR data(2000).
- [34] Liang, J., Zhou, J., Qian, Y., Wen, L., Bai, X. and Gao, Y. On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 55, 2 (2017), 862-880.
- [35] Mas, J.-F., Pérez-Vega, A., Ghilardi, A., Martínez, S., Loya-Carrillo, J. O. and Vega, E. A suite of tools for assessing thematic map accuracy. Geography Journal, 2014(2014).
- [36] Petersson, H., Gustafsson, D. and Bergstrom, D. Hyperspectral image analysis using deep learning—a review. IEEE, City, 2016.
- [37] Pan, B., Shi, Z. and Xu, X. R-VCANet: a new deep-learning-based hyperspectral image classification method. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 5 (2017), 1975-1986.
- [38] Roedl, G., Elmes, G. A. and Conley, J. Spatial technology applications. Springer, City, 2014.
- [39] Iordache, M.-D., Bioucas-Dias, J. M. and Plaza, A. Total variation spatial regularization for sparse hyperspectral unmixing. IEEE Transactions on Geoscience and Remote Sensing, 50, 11 (2012), 4484-4502.
- [40] Du, P., Xue, Z., Li, J. and Plaza, A. Learning discriminative sparse representations for hyperspectral image classification. IEEE Journal of Selected Topics in Signal Processing, 9, 6 (2015), 1089-1104.
- [41] Charles, A. S., Olshausen, B. A. and Rozell, C. J. Learning sparse codes for hyperspectral imagery. IEEE Journal of Selected Topics in Signal Processing, 5, 5 (2011), 963.
- [42] Yang, H., Du, Q. and Ma, B. Decision fusion on supervised and unsupervised classifiers for hyperspectral imagery. IEEE Geoscience and Remote Sensing Letters, 7, 4 (2010), 875-879.
- [43] Tarabalka, Y., Benediktsson, J. A. and Chanussot, J. Spectral-spatial classification of hyperspectral imagery based on partitioned clustering techniques. IEEE Transactions on Geoscience and Remote Sensing, 47, 8 (2009), 2973-2987.
- [44] Tarabalka, Y., Haavardsholm, T. V., Käsen, I. and Skauli, T. Real-time anomaly detection in hyperspectral images using multivariate normal mixture models and GPU processing. Journal of Real-Time Image Processing, 4, 3 (2009), 287-300.
- [45] Nascimento, J. M. and Dias, J. M. Does independent component analysis play a role in unmixing hyperspectral data? IEEE Transactions on Geoscience and Remote Sensing, 43, 1 (2005), 175-187.
- [46] Plaza, J., Plaza, A. J. and Barra, C. Multi-channel morphological profiles for classification of hyperspectral images using support vector machines. Sensors, 9, 1 (2009), 196-218.
- [47] Shaw, G. and Manolakis, D. Signal processing for hyperspectral image exploitation. IEEE Signal processing magazine, 19, 1 (2002), 12-16.
- [48] Merentitis, A., Debes, C. and Heremans, R. Ensemble learning in hyperspectral image classification: Toward selecting a favorable bias-variance tradeoff. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7, 4 (2014), 1089-1102.
- [49] Rajan, S., Ghosh, J. and Crawford, M. M. Exploiting class hierarchies for knowledge transfer in hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 44, 11 (2006), 3408-3417.
- [50] Peerbhay, K. Y., Mutanga, O. and Ismail, R. Random forests unsupervised classification: The detection and mapping of solanummauritanum infestations in plantation forestry using hyperspectral data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8, 6 (2015), 3107-3122.
- [51] Zhao, J., Zhong, Y., Wu, Y., Zhang, L. and Shu, H. Sub-pixel mapping based on conditional random fields for hyperspectral remote sensing imagery. IEEE Journal of Selected Topics in Signal Processing, 9, 6 (2015), 1049-1060.
- [52] Altmann, Y., Pereyra, M. and McLaughlin, S. Bayesian nonlinear hyperspectral unmixing with spatial residual component analysis. IEEE Transactions on Computational Imaging, 1, 3 (2015), 174-185.
- [53] Villa, A., Chanussot, J., Benediktsson, J. A. and Jutten, C. Spectral unmixing for the classification of hyperspectral images at a finer spatial resolution. IEEE Journal of Selected Topics in Signal Processing, 5, 3 (2011), 521-533.
- [54] Gu, Y., Wang, S. and Jia, X. Spectral unmixing in multiple-kernel hilbert space for hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 51, 7 (2013), 3968-3981.
- [55] Benediktsson, J. A., Palmason, J. A. and Sveinsson, J. R. Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Transactions on Geoscience and Remote Sensing, 43, 3 (2005), 480-491.
- [56] Bazi, Y. and Melgani, F. Toward an optimal SVM classification system for hyperspectral remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 44, 11 (2006), 3374-3385.
- [57] Sakla, W., Chan, A., Ji, J. and Sakla, A. An SVDD-based algorithm for target detection in hyperspectral imagery. IEEE Geoscience and Remote Sensing Letters, 8, 2 (2011), 384-388.
- [58] Gurram, P. and Kwon, H. Support-vector-based hyperspectral anomaly detection using optimized kernel parameters. IEEE Geoscience and Remote Sensing Letters, 8, 6 (2011), 1060-1064.
- [59] Heinz, D. C. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 39, 3 (2001), 529-545.
- [60] Heylen, R., Scheunders, P., Rangarajan, A. and Gader, P. Nonlinear unmixing by using different metrics in a linear unmixing chain. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8, 6 (2015), 2655-2664.
- [61] Dórido, I., Li, J., Gamba, P. and Plaza, A. A new hybrid strategy combining semisupervised classification and unmixing of hyperspectral data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7, 8 (2014), 3619-3629.
- [62] Cheng, Q., Varshney, P. K. and Arora, M. K. Logistic regression for feature selection and soft classification of remote sensing data. IEEE Geoscience and Remote Sensing Letters, 3, 4 (2006), 491-494.
- [63] Persello, C. and Bruzzone, L. Active learning for domain adaptation in the supervised classification of remote sensing images. IEEE

- Transactions on Geoscience and Remote Sensing, 50, 11 2012), 4468-4483.
- [64] Manolakis, D., Marden, D. and Shaw, G. A. Hyperspectral image processing for automatic target detection applications. Lincoln laboratory journal, 14, 1 2003), 79-116.
- [65] Chang, C.-I. and Chiang, S.-S. Anomaly detection and classification for hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 40, 6 2002), 1314-1325.
- [66] Villa, A., Chanussot, J., Benediktsson, J. A., Jutten, C. and Dambreville, R. Unsupervised methods for the classification of hyperspectral images with low spatial resolution. Pattern Recognition, 46, 6 2013), 1556-1568.
- [67] Balas, C., Epitropou, G. and Pappas, C. Multi/hyper-spectral imaging. Handbook of Biomedical Optics 2011), 131-164.
- [68] Zhao, W., Guo, Z., Yue, J., Zhang, X. and Luo, L. On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery. International journal of Remote sensing, 36, 13 2015), 3368-3379.
- [69] Hu, W., Huang, Y., Wei, L., Zhang, F. and Li, H. Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors, 2015 2015).
- [70] Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F. and Fraundorfer, F. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geoscience and remote sensing magazine, 5, 4 2017), 8-36.
- [71] Chen, Y., Zhao, X. and Jia, X. Spectral-spatial classification of hyperspectral data based on deep belief network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8, 6 2015), 2381-2392.
- [72] Li, W., Wu, G. and Du, Q. Transferred Deep Learning for Anomaly Detection in Hyperspectral Imagery. IEEE Geosci. Remote Sensing Lett., 14, 5 2017), 597-601.
- [73] Sankaran, S., Mishra, A., Ehsani, R. and Davis, C. A review of advanced techniques for detecting plant diseases. Computers and electronics in agriculture, 72, 1 2010), 1-13.
- [74] Duvvuru, R., Rao, G. N., Bendalam, S., Gemechu, R., Chundi, A., Naidu, D. S., Rao, C. R. and Rao, P. J. A Case Study On Child Labour-GIS Approach.
- [75] Mendoza, F., Lu, R., Ariana, D., Cen, H. and Bailey, B. Integrated spectral and image analysis of hyperspectral scattering data for prediction of apple fruit firmness and soluble solids content. Postharvest Biology and Technology, 62, 2 2011), 149-160.
- [76] Huang, M. and Lu, R. Apple mealiness detection using hyperspectral scattering technique. Postharvest Biology and Technology, 58, 3 2010), 168-175.
- [77] Ariana, D. P. and Lu, R. Evaluation of internal defect and surface color of whole pickles using hyperspectral imaging. Journal of Food Engineering, 96, 4 2010), 583-590.
- [78] Liu, L., Ngadi, M., Prasher, S. and Gariépy, C. Objective determination of pork marbling scores using the wide line detector. Journal of Food Engineering, 110, 3 2012), 497-504.
- [79] Arngren, M., Hansen, P. W., Eriksen, B., Larsen, J. and Larsen, R. Analysis of pregerminated barley using hyperspectral image analysis. Journal of agricultural and food chemistry, 59, 21 2011), 11385-11394.
- [80] Khan, M. J., Khan, H. S., Yousaf, A., Khurshid, K. and Abbas, A. Modern trends in hyperspectral image analysis: a review. IEEE Access, 6 2018), 14118-14129.

A Review on Scream Classification for Situation Understanding

Saba Nazir, Muhammad Awais
Dept. of Software Engineering
Govt. College University
Faisalabad, Pakistan

Sheraz Malik
Dept. of Information Technology
Govt. College University
Faisalabad, Pakistan

Fatima Nazir
Dept. of Software Engineering
Govt. College University
Faisalabad, Pakistan

Abstract—In our living environment, a non-speech audio signal provides a significant evidence for situation awareness. It also compliments the information obtained from a video signal. In non-speech audio signals, screaming is one of the events in which the people like security guard, care taker and family members are particularly interested in terms of care and surveillance because screams are atomically considered as a sign of danger. Contrary to this concept, this review is particularly targeting automated acoustic systems using non-speech class of scream believing that the screams can further be classified into various classes like happiness, sadness, fear, danger, etc. Inspired by the prevalent scream audio detection and classification field, a taxonomy has been projected to highlight the target applications, significant sound features, classification techniques, and their impact on classification problems in last few decades. This review will assist the researchers for retrieving the most appropriate scream detection and classification technique and acoustic parameters for scream classification that can assist in understanding the vocalization condition of the speaker.

Keywords—Scream classification; scream detection; acoustic parameters; surveillance; security

I. INTRODUCTION

In the past few decades, there have been several efforts regarding the classification of the acoustic data into classes. The audio data is very informative and a rich source of extraction for the type of content involving content-based classification of the acoustic signals. Human beings use vocal tract for producing speech sounds such as talking, singing, crying, and laughing. These sounds are further classified as speech or non-speech vocalizations. Speech consists of voices that are in the form of sentences and can be understood using different Natural Language Processing (NLP) techniques. The non-speech sounds include laugh, sneeze, cough, snore, and scream. These non-speech vocalizations are sometimes segregated from speech signals to extract additional information about the context, situation, or emotional state of the speaker. Scream is a non-speech signal that is caused by a loud vocalization when air passes through vocal folds with greater force than regular vocalizations. Most often, a scream is a reflex action or a response from an unexpected situation and it is strongly associated with emotional behavior of the speaker. It can have many forms like a scream of joy, danger, pain, surprise, etc.

Scream sound event classification and detection has wide applications in science due to which it has gained significant

importance in literature. Many real-life acoustic systems use scream detection in the areas like speaker identification [1], Audio-Surveillance Systems [2] and Home applications [3]. These systems use the knowledge extracted from scream detection and classification for processing. In this field, the conjunction of time-frequency features and machine learning classifier have achieved recent developments. Different techniques and methodologies have been established to differentiate speech and non-speech sounds. These include Support Vector Machines [3], band-limited spectral entropy [4], Deep Neural Networks (DNN) [5], Hidden Markov Model (HMM), sound event partitioning [6] and modulation power spectrum [7].

Most works on scream detection and classification emphasize on some crucial acoustic events, none cover the overall state-of-the-art for scream classification and detection. The current work varies of all preceding efforts in terms of emphasis, correctness as well as suitability. The aim of this review is to highlight the scream classification concerns and challenges to analyze and classify the screams from a variety of perspectives. Additionally, a comparative study is hereby presented that is based on the problem domain, sound features, and classification techniques. By overviewing this review, one can easily determine the problem domains where to put the scream efforts, using best sound parameters and scream classification techniques for situation understanding.

This review is planned as follows. Section 2 covers the data collection techniques and research methodology. Section 3 contains an overview of different classes of problem domains, sound features, and classification techniques. Section 4 evaluates the various data classes and argues on the comparison and accuracy rates. Finally, Section 5 concludes the key points in this review.

II. DATA COLLECTION

A review of 30 different research articles that are associated with scream detection and classification in various environments is presented. Highly cited and credible publications are used from different digital libraries for obtaining the research source. A thorough analysis is performed on all the articles to make sure that the content is pertinent to the research interests. Those classification problems that have hindered the further development and exploration in screaming environments, are discussed.

TABLE I. SELECTED RESEARCH ARTICLES WITH PROBLEMS DESCRIBED

#	Name/ Ref	Year	Problem	Detection/ Classification
1.	A. Pillai et al.[8]	2018	Classifying violent extensive audios like music, speech, gunshots, and screams.	Detection
2.	J. H. L. Hansen et al.[1]	2017	Analyzing human screams for text-independent speaker identification	Detection
3.	N. Hayasaka et al. [4]	2017	Detection of human scream considering noise robustness	Detection
4.	S. Chung et al. [9]	2017	Detecting screams for social problems and violent crimes in public places	Detection
5.	S. Mun et al. [10]	2017	Classification of acoustic scene using screams	Detection
6.	L. Girin [5]	2016	Automating screams detection in subway trains	Detection
7.	Y. Li et al. [11]	2016	Automatically classifying audio events like glass breaking, gunshots, footsteps, and screams for surveillance.	Detection
8.	A. Sharma et al. [12]	2016	Scream and cry detection in urban environments	Detection
9.	L. H. Arnal et al.[7]	2015	Using acoustic analysis, psychophysical experiments, and neuroimaging to isolate screaming features, and track their processing in the human brains	Detection
10.	J. H. L. Hansen et al. [13]	2015	Robustly detecting screams in noisy areas using unsupervised learning algorithm.	Detection
11.	M. Z. Zaheer et al. [14]	2015	Scream detection for existing CCTV cameras for better surveillance.	Detection
12.	M. K. Nandwana et al. [15]	2014	Finding out the impact of screaming on the performance of text independent speaker recognition systems	Detection
13.	M. Vacher et al. [16]	2014	Sound classification for patients and elderly people hospitalized at home.	Detection
14.	M. Vacher et al. [17]	2014	Detection and classification of acoustic events in a noisy environment	Detection
15.	B. Lei et al. [18]	2014	Power-efficient sound-event detection.	Detection
16.	K. Kato [19]	2013	Clarifying audio features of the death growl as well as screaming voice.	Detection
17.	B. UzKent et al. [20]	2012	Classification of non-speech environmental sounds using new feature set.	Detection
18.	M. Mark et al. [21]	2012	Investigate the consumption of power for a sound-event classification system at different stages	Detection
19.	W. Huang et al. [3]	2010	Detection of human screams using analytic and statistical features as a method of classification.	Detection
20.	C. Chan et al. [22]	2010	Scream/Non-scream classification in an abnormal situation such as bank robbery	Detection
21.	W. Liao et al. [23]	2009	Analysing non-speech human sounds, like scream, laugh, snore and sneeze.	Detection
22.	A. Fleury et al. [24]	2008	Home based global speech and sound recognition system for surveillance.	Detection
23.	L. Gerosa et al. [2]	2007	audio-based surveillance system to detect anomalous acoustic events like screams or gunshots.in public.	Detection
24.	C. Zhang et al. [25]	2007	Collectively consider the five speech modes in maintaining speech system performance for coding, speech, and speaker recognition.	Detection
25.	A. Rabaoui et al. [26]	2007	Improving one-class SVM classifier for sounds classification.	Detection
26.	P. C. Schön et al. [27]	2004	Develop a system to record and monitor level of anxiety sounds/calls in pigs.	Detection/ Classification
27.	M. Vacher et al. [28]	2004	Scream detection using transient models to ensure short detection delay in medical telesurvey	Detection
28.	E. R. Siebert et al. [29]	2003	Analyzing the structure and context of chimpanzee screams.	Detection
29.	N. E. O. Connor et al. [30]	2002	Unusual scene and sound detection in web-cam images using unsupervised learning	Detection
30.	R. A. Breguet et al. [31]	2000	Automatic detection/recognition of impulsive sounds, such as human screams, glass breaks, gunshots, door slams or explosions.	Detection

TABLE II. DATA CLASSIFICATION

Sr.	Type	Class
1.	Problem Domain	Surveillance Speaker Identification Feature Enhancement
2.	Feature Extraction	Temporal Spectral Prosodic
3.	Classification Techniques	Supervised Learning Unsupervised Learning

The selected data has been divided into different categories to carve out possible alternatives in several directions. Table I presents literature works related to the scenarios using scream detection or classification as major. The problem of each research article is described along with its ability of detection or classification of screams. Most of the authors are focusing on using the scream detection in the surveillance systems as in a common understanding the screams are a source of danger. Other have focused on whether enhancing the sound features of the systems under study or the identifying the speakers by their vocal scream samples. Only one author has worked indirectly on scream classification along with detection for animal screams.

These research studies analyze and compare the crucial aspects of different scream detection and classification methods. The major concerning factor is the accuracy of detection and classification stages, while minimizing the error rates and choosing the best possible sound features. In this review the emphasis is on the aspects of proficiency and accuracy of scream classification techniques.

On the first glance of Table I, it is very unsure to find out the loose ends and research gaps for a researcher who is new to this field. For this reason, each source is separated in terms of its problem domain, sounds parameters, type of classification technique used, and the results obtained in each case. All these categories are later further divided into different classes for even a broader understanding. Furthermore, tables and graphs are used in each class to compare on source with the other to find out which domain, parameter or technique is the best one to work out in future.

III. DATA CLASSIFICATION

The process of organizing data into groups and categories for its most effective and efficient use is broadly defined as data classification. As described above the collected data samples from different sources are analyzed based on the parameters discussed in Table II.

A. Problem Domain

A problem domain is the area of knowledge or application that desires to be analyzed and examined to solve a problem. Converging on a problem domain is simply focusing at only the topics of a person's interest, and apart from everything else. Based on the observations from various research sources, the problem domain has been divided into three categories: i) Surveillance, ii) Speaker Identification, and iii) Acoustic

Features Enhancement. All these categories are discussed in detail below:

1) *Surveillance*: Surveillance means managing, protecting, influencing, or directing the people by monitoring the abnormal activities or changing information in their surroundings [32]. Surveillance systems enable the remote observation of prevalent society for public safety and integrity. These observations can be made by some electronic devices like audio/video recordings or phone calls. Sound based surveillance systems enables remote public protection by analyzing sound samples collected from the target location or the target person. Screams plays an intense role in analyzing the situation analysis for any signs of danger.

2) *Speaker Identification*: Speaker identification systems are used to identify a person from voice biometrics. These systems use those human voice features that differ in different individuals. Screams can be used very effectively for text-independent speaker identification.

3) *Acoustic Feature Enhancement*: Quite a large set of scream literature is based on the techniques that are used to improve the acoustic features enhancement of scream detection and classification. These techniques help in increasing the robustness of the detection and classification for several different kind of sound-based scream dependent systems.

B. Feature Extraction

While evaluating and characterizing the contents of an audio stream, feature extraction plays a vital role. To analyze the a scream audio stream, the first step is extracting the concerned acoustic features form the audio frames.

TABLE III. CATEGORIZATION OF AUDIO FEATURES FOR SCREAM DETECTION

Category	Types
Temporal	Zero Crossing Rate (ZCR)
	Short Term Energy
Spectral	Mel-frequency Cepstral Coefficients (MFCC)
	Centroid
	Roll off
	Flux
	Tilt
	Spectral Entropy
	Signal Bandwidth
	Sub-Band Energy Ratio
	Linear Prediction
Prosodic	Fundamental Frequency /Pitch
	Intensity /Loudness
	Duration /Rhythm
	Log Energy

Table III represents different kind of acoustic features including Temporal, Spectral and Prosodic. This categorization is performed on the basis of diverse behaviour of acoustic parameters. These features can be extracted from audio signals or easy adaptability, robustness again noise and implementation.

1) *Temporal*: In a sound signal the amplitude fluctuation with time (the waveform signal) is represented as Temporal or time amplitude features. These acoustic features can be straightly extracted from raw sound signals for which no prior data is required. Typical temporal cases include amplitude-based features, zero-crossing rate (ZCR), and power-based features. Such features usually recommend a simple tactic to examine acoustic signals.

2) *Spectral*: Spectral/Cepstral features are resulted from short-term spectral features. Audio signals mostly speech and non-speech, speaker and language recognition rely on Cepstral features. The computation of cepstral is composed of three processes namely Fourier transform, inverse Fourier transform and logarithm [33]. These processes allow the identification of the purification and base frequency and of the audio signal. The different variants of Spectral features include Mel-frequency Cepstral Coefficients, Spectral Centroid, Spectral Flux, Spectral Roll off, Spectral Tilt, Spectral Entropy, Signal Bandwidth, Sub-Band Energy Ratio, and Linear Prediction. Generally, the temporal features are necessarily combined with spectral features for in-depth audio analysis. Consequently, the computational complexity of spectral features is higher than that of temporal features.

3) *Prosodic*: In the context of human listeners, to specify information with semantic sense, prosodic/ perceptual frequency features are used. On the other hand, the prosodic features define auditory signals in terms of mathematical and physical properties. These features are ordered based on semantically eloquent characteristics of sounds. These aspects include loudness/intensity, fundamental frequency, and rhythm.

C. Scream Classification Techniques

Scream classification can be performed using traditional classification tactics. An example of such tactics includes manual classification done by human experts. The experience and skills of a good analyst make this method more reliable. Though, it is time intense and arduous in spite of the precise results. To diminish human interaction for automating the detection and classification process, two approaches are widely used and applied for scream detection and classification. These two classification approaches are supervised and unsupervised that are highlighted in Table IV along with their sub-techniques. The use of semi-supervised learning algorithms is nearly non-considerable in terms of scream classification and hereby not a part of this review.

TABLE IV. MACHINE LEARNING TECHNIQUES FOR SCREAM CLASSIFICATION

Category	Classification Techniques
Supervised Learning	K-nearest-neighbors (KNN)
	Neural Networks (RBF, MLP)
	Support Vector Machines (SVM)
	Bayesian Networks
	Linear Discriminants
	Rule-based Classifiers
Unsupervised Learning	Neural Networks
	Hidden Markov Models (HMM)
	Gaussian Mixture Models (GMM)
	Clustering

1) *Supervised Learning Algorithms*: Supervised learning algorithms are those that aim to discover a relationship between a given input/vector and the desired output/supervisory signal. Once it analyses and figures out an association, it produces a pattern/inferred function which can be used for mapping new examples.

Supervised learning is extensively used in scream audio event detection systems. These techniques include K-nearest neighbor (k-NN), linear discriminant analysis, Bayesian networks, support vector machine, and rule-based machine algorithms. The obvious description or specification of these algorithms is to train the behavioral models with labelled data. This method holds high demand on resource consumption.

a) Instance-Based or K-Nearest-Neighbors (KNN)

The k-nearest neighbor algorithm (KNN) is the simplest and most efficient non- parametric algorithm from the family of instance-based learning [34]. The output of this algorithm depends on whether it is used for regression or classification. K-NN is a robust method that is proficient enough for organizing and segmenting audio streams into music, speech, environmental sounds, and silence [35]. The author in [11] used KNN for scream classification. The classification is done based on majority of neighbors. The object is allocated to the class that is in its k nearest neighbors where k is a positive integer. The value of k=1 depicts that the object is allocated to the class of exactly that single nearest neighbor. Although KNN is quite easy to implement but this technique requires memory and computation complexities. To overcome this problem, [36] and many other techniques have been developed.

b) Neural Networks

The Artificial Neural Network (ANN) is a data processing computing system which is vaguely encouraged by the biological neural networks, such as the animal or human brain process information. For audio events the Radial Basis Function (RBF) and Multi-Layer Perceptron (MLP) were applied in Artificial Neural Networks (ANNs) for supervised audio classification to decrease misclassification errors. In MLP, input datasets are mapped onto appropriate output sets. The most common use of MLP is in automatic phoneme recognition tasks [37]. A particular case [38] of feed-forward network is Radial Basis Function (RBF) which creates a linear map from the hidden space to the output space.

c) Rule-Based Classifiers

A rule-based machine learner identifies and utilize a set of relational rules that cooperatively show the knowledge captured by the system. This contrasts with the other machine learners where a singular model is commonly identified that can be applied universally on any instance to make a prediction. A variation of this classifier is fuzzy rule-based classifier (FRBC) that is efficiently being used for numerous classification tasks. Auditory event detection in fuzzy set-oriented contains the information concerning to a set of rules that classify the several characteristics of the fuzzy rule base in the training data [39]. The disadvantage of fuzzy operators is that there is no specific way to define fuzzy operators especially symbolic variables. The classification problem of non-speech human voice was solved [40] using fuzzy integral and some of the associated fuzzy measures.

d) Bayesian Networks

A Bayesian/Bayes/Belief network is a graphical model that probabilistically signifies a set of variables and their inter dependencies using a directed acyclic graph (DAG). There are the variants of the Bayesian network include: 1) serial, 2) divergent, and 3) convergent. It does fast supervised classification due to which it is appropriate for forecasting and classification tasks on complex large-scale datasets. Various multi-modal [41]-[43] have been projected to resolve the glitches in acoustic and speech segmentation in movies or robot speech under noise conditions.

e) Linear Discriminants

Linear discriminant analysis (LDA) is used to find a linear combination of features that classifies two or more classes of objects or events. The resulting combination can then be used as a classifier, or for dimensionality reduction. LDA basically transfers raw data into a feature space [44] supporting a more robust classification.

f) Support Vector Machines (SVM)

Support vector machines (SVMs) are valuable machine learning method for complicated data classification problems [45]. A training set is provided to the SVM by a set called input vector. SVMs separate two types or classes by maximizing the margin between the class boundaries and the nearest sample to it.

2) *Unsupervised Learning Algorithms:* Unsupervised Learning algorithms are applied to infer a function or

conclusions from unlabeled input data. As the data is unlabeled so its process involves finding and correlating the labels. The main objective of unsupervised learning is to examine the information and discovering similarities between the objects.

In unsupervised learning, the most common method is Cluster analysis that utilizes heuristic data for analyzing and finding hidden classes and patterns in audio data. Similarity measurement is used in clustering that is based upon metrics like Euclidean distance and probabilistic distance [46]. Some common algorithm for clustering are: 1) Gaussian Mixture Models, 2) Clustering, 3) Hidden Markov Models, and 4) Neural Networks.

a) Gaussian Mixture Models (GMM)

Gaussian mixture models (GMMs) are unsupervised classification methods. These methods are extensively used in speech/voice recognition and sensing and hence can be applied to. GMM assumes that all the data points are created from a mixture that contains several Gaussian distributions with unidentified parameters.

b) Clustering

Hierarchical clustering (HC) also called hierarchical cluster analysis is a technique of cluster analysis that is aimed at building a hierarchy of clusters by recursively merging or dividing the patterns [47], [48]. It uses two kinds of strategies. One includes constructing a hierarchy from the bottom up (agglomerative) after calculating the similarities among all pairs of clusters iteratively merging the most similar pair. The other top down (Divisive) approach performs splits recursively moving down the hierarchy.

In Partitioning approaches, samples are repositioned by transferring from one cluster to the other. This method initially requires the total number of clusters that will be pre-set by the user. The well-known methods in this field include K-means and its variants [48], [49].

c) Hidden Markov Models (HMM)

A hidden Markov model (HMM) is based on unobserved or hidden variable states. This model is a statistical Markov chain. The unobserved states are obtained based on a particular emission function that is resultant of some observable symbols [50]. The hidden Markov model can be considered to be the simplest dynamic Bayesian network. C. Chan et al. [22], M. Vacher et al. [16] used HMM for scream classification.

d) Neural Networks

Artificial neural networks (ANNs) are huge computing systems working together, consisting huge number of processors and their interconnections. The ANNs can solve reliable and efficient classification problems obtaining high tolerance and adaptability [51]. The most commonly used neural network models for unsupervised learning algorithms are Self organizing Map (SOM) and Adaptive Resonance Theory (ART).

IV. RESULTS AND DISCUSSION

A total of 30 research articles based on scream classification and detection are used and compared based on

problem domains, sound features, and classification techniques. A quick analysis of the review for each case is presented below:

A. Analysis of Problem Domain

Three main problem domains for scream classification include Surveillance, Speaker Identification and Feature Enhancement. The relevant research articles are separated for each problem domain. The division of articles is hereby shown in Table V. It represents that out of 30 articles, 19 belong to individual person or public surveillance, 3 belong to the identification of the speaker and 8 discussed the methods and mechanisms to enhance and enrich the scream sound vocal experimental results. In the next step the overall percentages are calculated for these problem domains to find out which one is lagging and needs further exploration (Fig. 1).

TABLE V. DIVISION OF PROBLEM DOMAINS BASED ON LITERATURE

Sr.	Problem Domain	References	Total
1.	Surveillance	S. Chung et al. [9], S. Mun et al. [10], L. Girin [5], Y. Li et al. [11], [12], L. H. Arnal et al.[7], M. Z. Zaheer et al. [14], M. Vacher et al. [16], M. Vacher et al. [17], B. Uz Kent et al. [20], M. Mark et al. [21], W. Huang et al. [3], C. Chan et al. [22], A. Fleury et al. [24], L. Gerosa et al. [2], P. C. Schön et al. [27], M. Vacher et al. [28], E. R. Siebert et al. [29], N. E. O. Connor et al. [30]	19
2.	Speaker Identification	J. H. L. Hansen et al.[1], M. K. Nandwana et al. [15], C. Zhang et al. [25]	3
3.	Feature Enhancement	A. Pillai et al.[8], N. Hayasaka et al. [4], J. H. L. Hansen et al. [13], B. Lei et al. [18], K. Kato [19], W. Liao et al. [23], A. Rabaoui et al. [26], R. A. Breguet et al. [31]	8

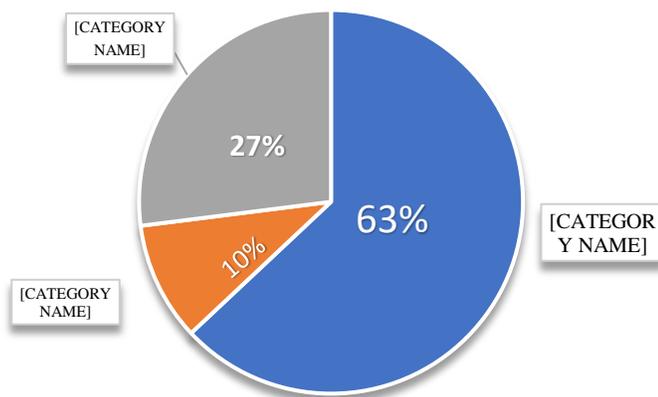


Fig. 1. Percentage usage of scream detection in various problem domains.

With the increasing rate of public crime occurrences (like on streets and transports), and danger to the precious human lives, surveillance systems based on audio analysis of screams are rapidly becoming popular. This is because the screams are usually considered and interpreted as to be the signals of survival in humans. Such systems can help majorly in medical surveys, audio scene classification, embedded transport environments like buses and trains, and 24x7 monitoring for the signs of distress in humans' daily routine.

Fig. 1 indicates that the Surveillance domain is more enriched with scream detection and classification because of the two reasons 1) Increasing number of health and safety issues and, 2) Screams are a sign of danger.

B. Analysis of Scream Sound Features

It is computationally expensive to utilize all the sound features for scream classification, so it is a common practice to mix-up one or two type of features together to achieve the best results in conjunction with classification techniques.

TABLE VI. TAXONOMY OF SCREAM FEATURE TYPES

Sr.	Feature Type	References
1.	Spectral (S)	L. Girin [5], S. Chung et al. [9], S. Mun et al. [10], Y. Li et al. [11], A. Sharma et al. [12], M. K. Nandwana et al. [15], M. Vacher et al. [16], B. Lei et al. [18], A. Fleury et al. [24], A. Rabaoui et al. [26], P. C. Schön et al. [27], R. A. Breguet et al. [31]
2.	Prosodic (P)	K. Kato [19]
3.	Temporal, Spectral (TP)	J. H. L. Hansen et al.[1], A. Pillai et al.[8], M. Vacher et al. [17], M. Vacher et al. [28], E. R. Siebert et al. [29], N. E. O. Connor et al. [30]
4.	Spectral, Prosodic (SP)	W. Huang et al. [3], N. Hayasaka et al. [4], M. Z. Zaheer et al. [14], B. Uz Kent et al. [20], W. Liao et al. [23], C. Zhang et al. [25]
5.	Temporal, Spectral, Prosodic (TSP)	L. Gerosa et al. [2], L. H. Arnal et al.[7], J. H. L. Hansen et al. [13], M. Mark et al. [21], C. Chan et al. [22]

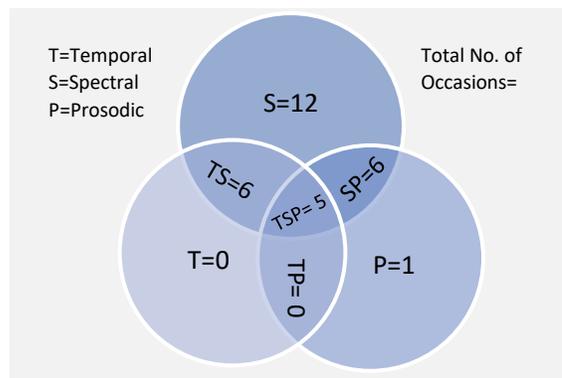


Fig. 2. Basic Venn diagram for the use of sound parameter on several occasions.

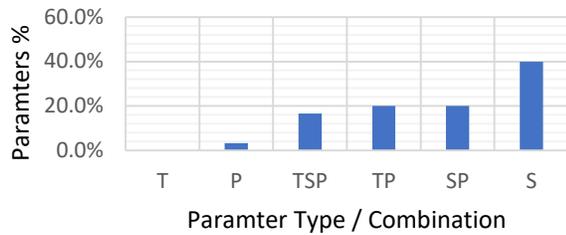


Fig. 3. Percentage usage of individual and combined parameter.

While exploring the sound features it can be observed that some of the articles are using the combined feature approach. Following this, a taxonomy has been developed (described in Table VI). The temporal features cannot be effectively used separately so no article has independently used these features but in combination with other types.

Spectral and Prosodic features are used independently as well as in combination. Table VI describes all of the articles under consideration and the type of sound features they have used or recommended for scream classification. The results of this step are shown in Fig. 2 and 3.

In Fig. 2, S=Spectral, P=Prosodic, T=Temporal, TS=Temporal and Spectral, SP= Spectral and Prosodic, TP= Temporal and Prosodic and TSP= Temporal, Spectral and Prosodic. It also shows that the most commonly used sound features are spectral. Out of 30 researches, 12 used spectral features independently. The second-best features are the combination of either TS or SP. While no one recommended T or TP.

The results are presented by calculating the percentages for each type or combination. The percentage evaluation is shown in Fig. 3 which clearly expresses that the spectral parameters are the most recommended ones to achieve the best scream classification with 40% of usability.

Further we see that there are further many forms of each category of scream sound feature. Table VII describes all the considered scream articles with the type of sound feature they have used in detail.

In the last step, it has been concluded that spectral features are highly recommended in literature for scream classification. The basic purpose of this step is to figure out that out of many forms of Spectral features which one shows the best performance out of all.

Fourier transform is used to convert time-domain signal into frequency domain for obtaining spectral features. These features are quite helpful in identifying the notes, pitch, rhythms and melody.

The results of this step are shown in Fig. 4. It can be clearly observed that Mel-frequency Cepstral Coefficients are the most used and highly recommended sound feature for scream classification. It can either be used individually or in combination with other sound features. MFCC are extensively applied in voice recognition because of the reason that these features are very similar to human listening. In more complicated and complex signals such as speech or music where the signal changes its properties over time, it is evidently more meaningful to refer to the altering frequency content over a smaller time interval than an infinite time interval.

TABLE VII. AUDIO FEATURE CATEGORIZATION FOR SCREAM DETECTION AND CLASSIFICATION

Category	Types	References
Temporal	Zero Crossing Rate (ZCR)	A. Pillai et al.[8], M. Vacher et al. [17], M. Vacher et al. [28], C. Chan et al. [22], L. Gerosa et al. [2]
	Short Term Energy	M. Mark et al. [21], A. Pillai et al.[8], J. H. L. Hansen et al.[1], L. H. Arnal et al.[7], J. H. L. Hansen et al. [13], E. R. Siebert et al. [29], N. E. O. Connor et al. [30]
Spectral	Mel-frequency Cepstral Coefficients (MFCC)	M. Mark et al. [21], J. H. L. Hansen et al.[1], N. Hayasaka et al. [4], L. H. Arnal et al.[7], M. Vacher et al. [28], S. Chung et al. [9], L. Gerosa et al. [2], S. Mun et al. [10], L. Girin [5], Y. Li et al. [11], A. Sharma et al. [12], J. H. L. Hansen et al. [13], M. Vacher et al. [16], B. Lei et al. [18], W. Huang et al. [3], A. Fleury et al. [24], A. Rabaoui et al. [26], M. K. Nandwana et al. [15], B. Uzkent et al. [20], W. Liao et al. [23], C. Zhang et al. [25]
	Spectral Centroid	M. Mark et al. [21], A. Pillai et al.[8], M. Vacher et al. [17], M. Vacher et al. [28], L. Gerosa et al. [2], R. A. Breguet et al. [31], W. Liao et al. [23],
	Spectral Roll off	A. Pillai et al.[8], M. Vacher et al. [17], M. Vacher et al. [28], L. Gerosa et al. [2], W. Liao et al. [23]
	Spectral Flux	E. R. Siebert et al. [29], L. Gerosa et al. [2], M. Z. Zaheer et al. [14], R. A. Breguet et al. [31]
	Spectral Tilt	L. Gerosa et al. [2], R. A. Breguet et al. [31], C. Zhang et al. [25]
	Spectral Entropy	M. Mark et al. [21], A. Pillai et al.[8], N. Hayasaka et al. [4], W. Liao et al. [23]
	Signal Bandwidth	M. Mark et al. [21], W. Liao et al. [23]
	Sub-Band Energy Ratio	J. H. L. Hansen et al.[1], C. Chan et al. [22], M. Z. Zaheer et al. [14], C. Zhang et al. [25]
Prosodic	Linear Prediction	P. C. Schön et al. [27], N. E. O. Connor et al. [30]
	Pitch/Fundamental Frequency	M. Mark et al. [21], L. H. Arnal et al.[7], C. Chan et al. [22], L. Gerosa et al. [2], J. H. L. Hansen et al. [13], M. Z. Zaheer et al. [14], K. Kato [19], B. Uzkent et al. [20], W. Liao et al. [23]
	Loudness/Intensity	L. Gerosa et al. [2], K. Kato [19], C. Zhang et al. [25]
	Rhythm/Duration	C. Chan et al. [22], K. Kato [19], C. Zhang et al. [25]
	Log Energy	N. Hayasaka et al. [4], W. Huang et al. [3]

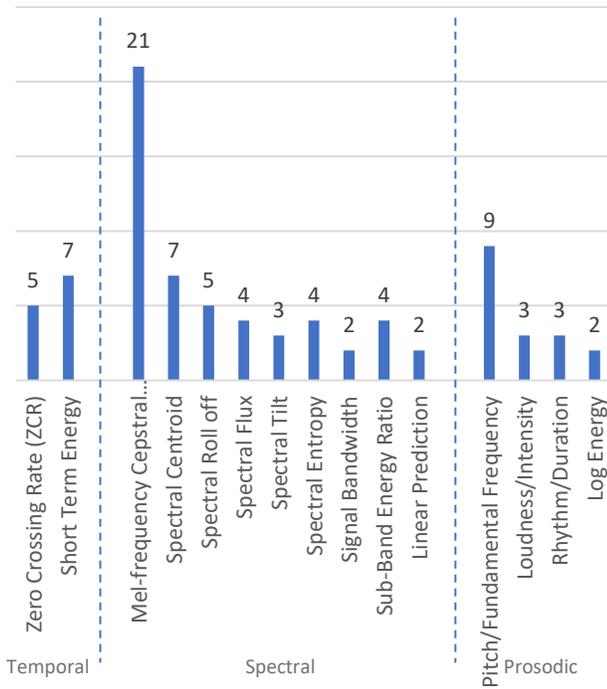


Fig. 4. Division of articles based on sound features.

C. Analysis of Classification Techniques

There are two clear divisions of sound event detection approaches: supervised, unsupervised or combination. These approaches are studied perceptibly however still suffer from a scarcity of additional thorough and complete analysis on classification approaches, primarily in scream signal classification. This review documents the scream classification with two subclasses in conjunction with a close review of every class.

This taxonomy has been shown in Table VIII. The referenced articles in each category are carefully observed and assigned to the relevant class. Some of the techniques are using supervised and unsupervised approach independently while the

others are using a combination of both approaches (separately). This table is not for comparison as the datasets and the sound features are used differently. It is just providing a review of the current illustrative approaches.

A more precise view is presented in Fig. 5, where 11 researches used supervised, 13 used un-supervised and 4 used combined scream classification approaches. Furthermore, the generic analytical view of classification approach is shown in Fig. 6, where the percentage calculations are performed in each case. It can clearly be seen that the un-supervised approaches have been more successfully been applied than other approaches in the last 18 years for scream detection and classification. For this purpose, the supervised and unsupervised scream classification techniques are further explained and analyzed in the next section.

1) *Supervised Learning Algorithms*: Supervised learning algorithms are categorized as K-nearest-neighbors (KNN) or instance-based, neural networks, rule-based Classifiers, linear discriminant, Bayesian networks, and support vector machines (SVM).

The primary purpose of this review is to present supervised learning approaches based on scream classification. The future researchers can find out the ways to explore the automated acoustic environments and systems. The most recent experimental research works related to screams classifications and detection are summarized in Table IX. It presents the latest methods for undertaking scream classification and detection issues based on supervised learning methods.

Accuracies of classifiers are statistically compared and calculated by finding out the total no. of researches along with their classification results. By finding the individual accuracy of each supervised learning classification technique mentioned in the literature, average accuracies have been calculated to find out which techniques is providing the best results.

Fig. 7 shows the percentage accuracies of each technique along with the authors and references independently. N. Hayasaka et al. [4] is leading while using SVMs with accuracy rate of 94.6%.

TABLE VIII. SCREAM CLASSIFICATION TECHNIQUES

Sr.	Feature Type	References
1.	Supervised	W. Huang et al. [3], L. Girin [5], L. H. Arnal et al.[7], A. Pillai et al.[8], A. Sharma et al. [12], B. Lei et al. [18], B. Uz Kent et al. [20], M. Mark et al. [21], W. Liao et al. [23], A. Rabaoui et al. [26], P. C. Schön et al. [27],
2.	Un- Supervised	J. H. L. Hansen et al.[1], L. Gerosa et al. [2], S. Mun et al. [10], J. H. L. Hansen et al. [13], M. Z. Zaheer et al. [14], M. K. Nandwana et al. [15], M. Vacher et al. [16], M. Vacher et al. [17], C. Chan et al. [22], A. Fleury et al. [24], C. Zhang et al. [25], N. E. O. Connor et al. [30], R. A. Breguet et al. [31]
3.	Both	N. Hayasaka et al. [4], S. Chung et al. [9], Y. Li et al. [11], M. Vacher et al. [28]

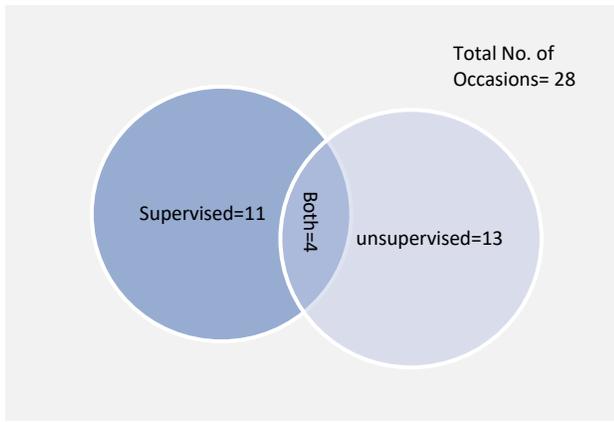


Fig. 5. Representation of scream classification techniques.

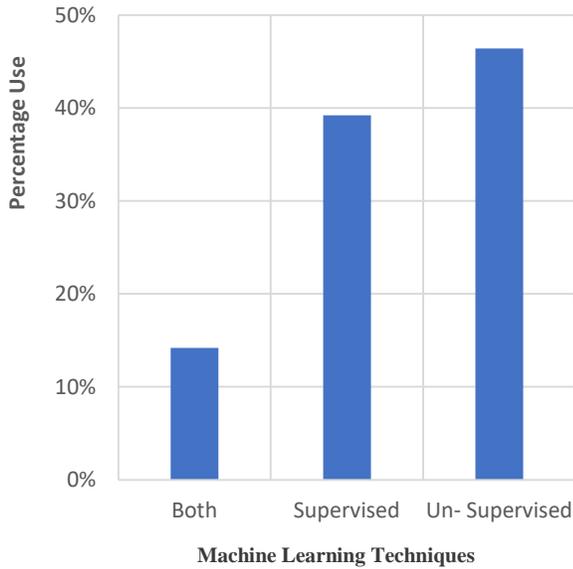


Fig. 6. Percentage usage of machine learning techniques.

Fig. 8 shows the average accuracies of all supervised scream classification techniques. It can be clearly observed that Linear discriminants are producing the highest accuracy rate of 96.1% and after that the KNN with accuracy rate of 94%.

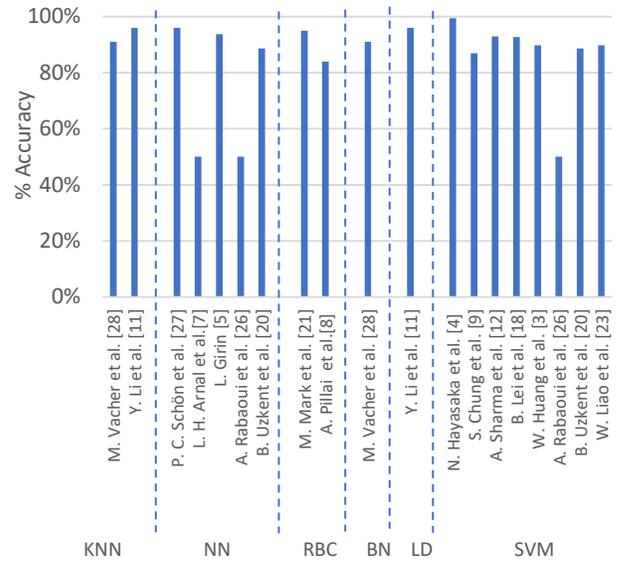


Fig. 7. Accuracies of supervised scream classifiers.

2) *Unsupervised Learning Algorithms*: Unsupervised learning algorithms comprehend a major learning paradigm and have drawn considerable attention in past few decades, as shown by the growing range of research publications in this field. The unsupervised methods for scream detection and classification are classified into four classes: Clustering, GMM, HMM and NN.

Table X lists the most significant research works and their average accuracies dealing with scream detection and classification problems associated with unsupervised approaches to present some solutions to the problems restraining the performance of scream classification systems for situation understanding.

Fig. 9 shows the percentage accuracies of each technique along with the authors and references independently. M.Z. Zaheer et al. [14] achieved 100% scream detection accuracy with GMM technique. Another classification technique used by N. Hayasaka et al. [4] achieved an accuracy rate of 99% again with GMM.

TABLE IX. SCREAM CLASSIFICATION TECHNIQUES

Category	Classification Techniques	References	Total Articles	Average Accuracies
Supervised Learning	Instance-based or K-nearest-neighbors (KNN)	M. Vacher et al. [28], Y. Li et al. [11]	2	94%
	Neural Networks (RBF, MLP)	P. C. Schön et al. [27], L. H. Arnal et al. [7], L. Girin [5], A. Rabaoui et al. [26], B. Uz Kent et al. [20]	5	76%
	Rule-based Classifiers	M. Mark et al. [21], A. Pillai et al. [8]	2	90%
	Bayesian Networks	M. Vacher et al. [28]	1	91%
	Linear Discriminants	Y. Li et al. [11]	1	96.1%
	Support Vector Machines (SVM)	N. Hayasaka et al. [4], S. Chung et al. [9], A. Sharma et al. [12], B. Lei et al. [18], W. Huang et al. [3], A. Rabaoui et al. [26], B. Uz Kent et al. [20], W. Liao et al. [23]	8	86.3%

TABLE X. UNSUPERVISED LEARNING TECHNIQUES FOR SCREAM CLASSIFICATION

Category	Classification Techniques	References	Total Articles	Average Accuracies
Unsupervised Learning Algorithms	Hierarchical and Partition Clustering	N. E. O. Connor et al.[30], M. K. Nandwana et al. [15],	2	76%
	Gaussian Mixture Models (GMM)	J. H. L. Hansen et al.[1], N. Hayasaka et al. [4], M. Vacher et al. [17], M. Vacher et al. [28], S.Chung et al. [9], L. Gerosa et al. [2], S. Mun et al. [10], Y. Li et al. [11], M. Z. Zaheer et al. [14], M. Vacher et al. [16], A. Fleury et al. [24], R. A. Breguet et al. [31], C. Zhang et al. [25]	13	86%
	Hidden Markov Models (HMM)	C. Chan et al. [22], M. Vacher et al. [16], A. Fleury et al. [24], R. A. Breguet et al. [31]	4	75%
	Neural networks (Self-organizing map)	S. Mun et al. [10], J. H. L. Hansen et al. [13]	2	83%

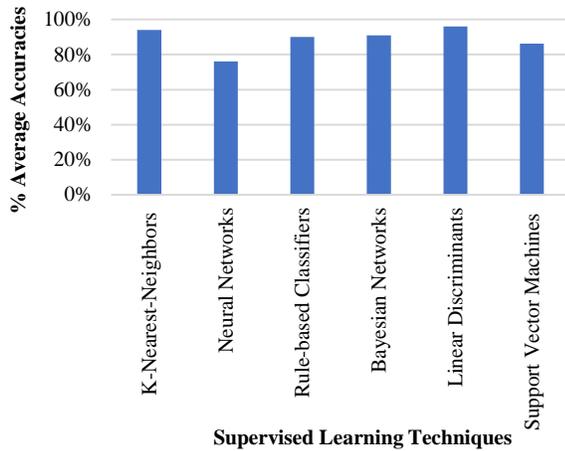


Fig. 8. Average accuracies of supervised stream classifiers.

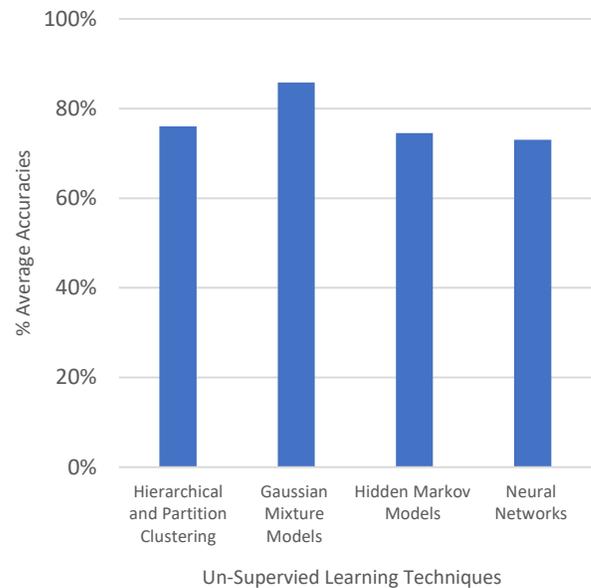


Fig. 10. Average accuracies of un-supervised stream classifiers.

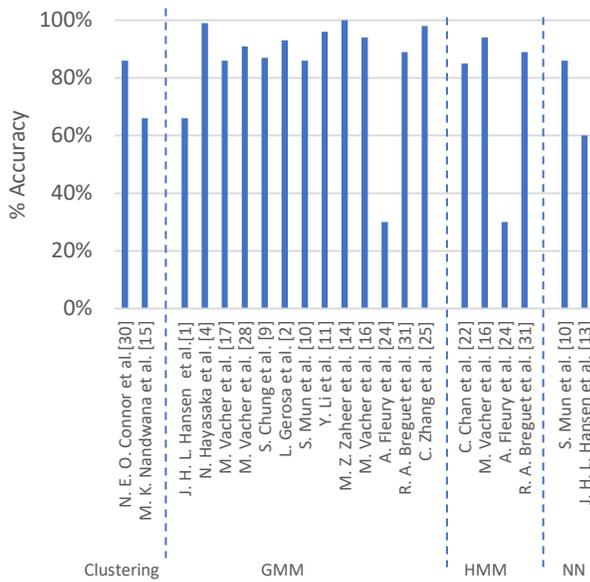


Fig. 9. Accuracies of unsupervised stream classifiers.

The overall average accuracies of the four un-supervised stream classifiers are calculated and plotted in Fig. 10. It can clearly be observed that GMMs are producing the best results with an average classification accuracy rate of 86%.

a) Combining Results

The results of overall review are converged in Table XI. It shows all of the research articles of last 18 years (from 2000-2018) based on the specified sound parameters and classification techniques. The accuracy percentage and the effective error rate (ERR) for each article is also mentioned. Fig. 11 plots these results for scream detection and scream sound classification.

Maximum percentage accuracies are used. Some of the researchers have done descriptive studies so that if the numerical results are not provided then these are supposed to be 50% accurate to show that the results are encouraging.

TABLE XI. BRIEF REVIEW OF SCREAM LITERATURE

#	Name/ Ref.	Year	Problem Domain			Sound Parameters			Classification Technique/es		Detection/ Classification		Results	
			Surveillance	Speaker Identification	Feature Enhancement	Temporal	Spectral	Prosodic	Supervised	Un-Supervised	Detection	Classification	% Accuracy	% EER
1.	A. Pillai et al.[8]	2018			✓	✓	✓		✓		✓		84%.	16%
2.	J. H. L. Hansen et al.[1]	2017		✓		✓	✓			✓	✓		66.67 %	33.33%
3.	N. Hayasaka et al. [4]	2017			✓		✓	✓	✓	✓	✓		99.45%	0.55%
4.	S. Chung et al. [9]	2017	✓				✓		✓	✓	✓		87.035%	12.965%
5.	S. Mun et al. [10]	2017	✓				✓			✓	✓		86.3%.	13.7%
6.	L. Girin [5]	2016	✓				✓		✓		✓		93.8%	6.2%
7.	Y. Li et al. [11]	2016	✓				✓		✓	✓	✓		96.1%	3.9%
8.	A. Sharma et al. [12]	2016	✓				✓		✓		✓		93.16%	6.84%
9.	L. H. Arnal et al.[7]	2015	✓			✓	✓	✓	✓		✓		50%	50%
10.	J. H. L. Hansen et al. [13]	2015			✓	✓	✓	✓	✓		✓		60%	40%
11.	M. Z. Zaheer et al. [14]	2015	✓				✓	✓		✓	✓		100%	0%
12.	M. K. Nandwana et al. [15]	2014		✓			✓			✓	✓		66.67%	33.33%
13.	M. Vacher et al. [16]	2014	✓				✓			✓	✓		94%	06%
14.	M. Vacher et al. [17]	2014	✓			✓	✓			✓	✓		86.46%	13.54%
15.	B. Lei et al. [18]	2014			✓		✓		✓		✓		92.76%	7.24%
16.	K. Kato [19]	2013			✓			✓	N/A	N/A	✓		50%	50%
17.	B. Uz kent et al. [20]	2012	✓				✓	✓	✓		✓		88.7%	11.3%
18.	M. Mark et al. [21]	2012	✓			✓	✓	✓	✓		✓		95%	05%
19.	W. Huang et al. [3]	2010	✓				✓	✓	✓		✓		89.815%	10.185%
20.	C. Chan et al. [22]	2010	✓			✓	✓	✓		✓	✓		85.75%	14.25%
21.	W. Liao et al. [23]	2009			✓		✓	✓	✓		✓		89.81%	10.19%
22.	A. Fleury et al. [24]	2008	✓				✓			✓	✓		30.43%	69.57%
23.	L. Gerosa et al. [2]	2007	✓			✓	✓	✓		✓	✓		93%,	07%
24.	C. Zhang et al. [25]	2007		✓			✓	✓		✓	✓		98.5%	1.5%
25.	A. Rabaoui et al. [26]	2007			✓		✓		✓		✓		50%	50%
26.	P. C. Schön et al. [27]	2004	✓				✓		✓		✓	✓	96%	04%
27.	M. Vacher et al. [28]	2004	✓			✓	✓		✓	✓	✓		91%	09%
28.	E. R. Siebert et al. [29]	2003	✓			✓	✓			N/A	✓		50%	50%
29..	N. E. O. Connor et al. [30]	2002	✓			✓	✓			✓	✓		86.67%	13.33%
30.	R. A. Breguet et al. [31]	2000			✓		✓			✓	✓		89%	11%

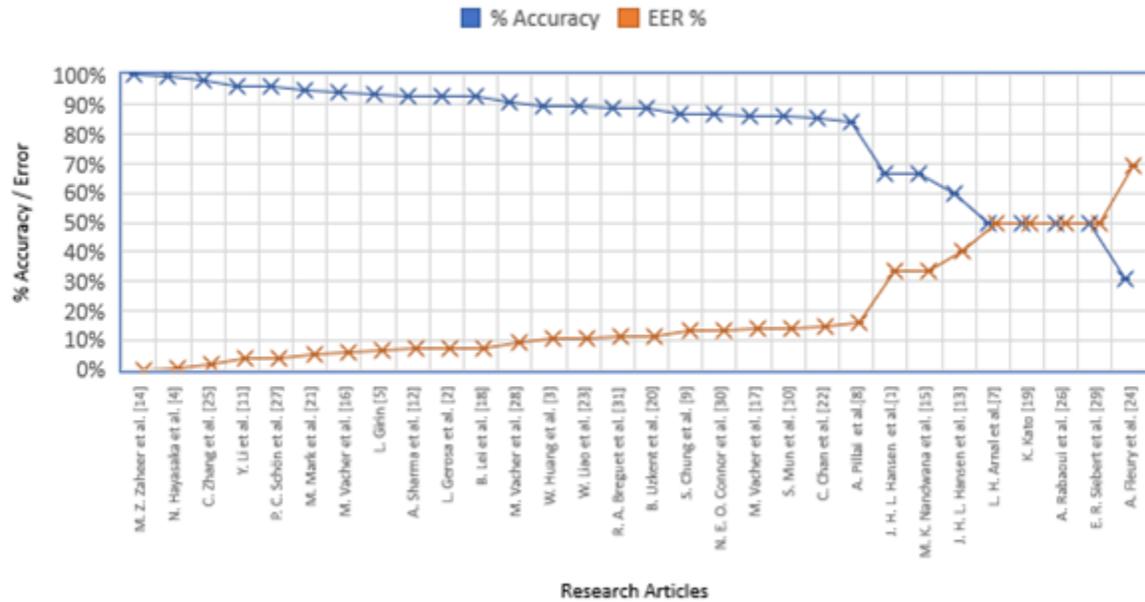


Fig. 11. Accuracy and EER for scream detection and classification.

It can be clearly observed that only a single research conducted by P. C. Schön et al. [27] in 2004 has focused on scream detection as well as classification. But this research is based on chimpanzee screams. The authors have figured out the ways in which the chimpanzees can be understood and what different kind of meanings can be driven from their screams. Two of the researches i.e. K. Kato [19] and E. R. Siebert et al. [29], have not used machine learning techniques instead they have developed their own for scream detection and classification.

So, there is a clear and a wide scope for scream classification non-understanding the situations in which they occur and to support the embedded sound-based systems especially surveillance systems to make the humans and animals out of danger.

V. CONCLUSION

A thorough analysis is presented on researchers' attempts related to scream detection and classification techniques. An in-depth taxonomy of scream detection and classification systems was presented in this review. The concerning efforts are expected to maximize scream signal detection and classification accuracy and understanding the surrounding situation of a speaker. The focus of this review is on machine learning and classification methods as well as essential sound parameters for scream-based audio embedded systems.

Although the best combination that can be concluded is that for the case of scream classification, unsupervised learning technique i.e. GMM can be applied using spectral sound features necessarily including MFCC in the field of surveillance. Because in surveillance scream detection and sound classification has been implemented in remarkably high percentage, so there are chances that the surveillance systems based on scream detection and classification, are causing a higher risk to humanity. But these results are concluded on the

information and statistics that is based on different kind of data sets using various combinations of sound parameters and classification techniques. The results may vary based on the datasets used and the background noise level.

In future, this review can be beneficial for the researchers to conduct a mechanism for scream classification and to understand the best possible alternatives in terms of classification techniques and sound parameters. A system can be developed using the concluded research to find out the differences in different classes of screams like joy, fear, sadness, etc. and to find out that how such kind of research can be helpful for understanding the surroundings of a speaker.

REFERENCES

- [1] J. H. L. Hansen, M. K. Nandwana, and N. Shokouhi, "Analysis of human scream and its impact on text-independent," vol. 2957, 2017.
- [2] L. Gerosa, M. Tagliasacchi, F. Antonacci, and I. Politecnico, "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems," 2007.
- [3] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream Detection for Home Applications," pp. 15–18, 2010.
- [4] N. Hayasaka, A. Kawamura, and N. Sasaoka, "Noise-robust scream detection using band-limited spectral entropy," *AEUE - Int. J. Electron. Commun.*, vol. 76, pp. 117–124, 2017.
- [5] L. Girin, "Deep Neural Networks For Automatic Detection Of Screams And Shouted Speech In Subway Trains Ifsttar , COSYS , LEOST , Villeneuve d ' Ascq INRIA Grenoble Rh'," pp. 6460–6464, 2016.
- [6] B. Lei and M. Mak, "Robust scream sound detection via sound event partitioning," 2015.
- [7] L. H. Arnal et al., "Human Screams Occupy a Privileged Niche in the Communication Soundscape Report Human Screams Occupy a Privileged Niche in the Communication Soundscape," *Curr. Biol.*, pp. 1–6, 2015.
- [8] A. Pillai and P. Kaushik, "AC : An Audio Classifier to Classify Violent Extensive Audios," 2018.
- [9] S. Chung and Y. Chung, "Scream sound detection based on SVM and GMM," no. 1, pp. 1–4, 2017.
- [10] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, "Deep Neural Network Based Learning and Transferring Mid-Level Audio Features

- for Acoustic Scene Classification,” pp. 796–800, 2017.
- [11] Y. Li and G. Liu, “Sound Classification Based On Spectrogram,” no. 2, 2016.
- [12] A. Sharma and S. Kaul, “Two-Stage supervised learning-based method to detect screams and cries in urban environments,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 2, pp. 290–299, 2016.
- [13] J. H. L. H. Mahesh Kumar Nandwana, Ali Ziayi, “Robust Unsupervised Detection Of Human Screams In Noisy Acoustic Environments,” pp. 161–165, 2015.
- [14] M. Z. Zaheer, J. Y. Kim, H. G. Kim, and S. Y. Na, “A preliminary study on deep-learning based screaming sound detection,” 2015 5th Int. Conf. IT Converg. Secur. ICITCS 2015 - Proc., 2015.
- [15] M. K. Nandwana, J. H. L. Hansen, E. Jonsson, and C. Science, “Analysis and Identification of Human Scream: Implications for Speaker Recognition,” no. September, pp. 2253–2257, 2014.
- [16] M. Vacher and M. Vacher, “Sound Classification in a Smart Room Environment : an Approach using GMM and HMM Methods,” 2014.
- [17] M. Vacher et al., “Sound Detection and Classification for Medical Telesurvey,” 2014.
- [18] B. Lei, “Sound-Event Partitioning and Feature Normalization for Robust Sound-Event Detection,” no. August, pp. 389–394, 2014.
- [19] K. Kato and A. Ito, “Acoustic features and auditory impressions of death growl and screaming voice,” *Proc. - 2013 9th Int. Conf. Intell. Inf. Hiding Multimed. Signal Process. IIH-MSP 2013*, pp. 460–463, 2013.
- [20] B. Uz Kent, B. D. Barkana, and H. Cevikalp, “Non-speech environmental sound classification using SVMs with a new set of features,” no. January 2015, 2012.
- [21] M. Mak and S.-Y. Kung, “Low-Power Svm Classifiers For Sound Event Classification On Mobile Devices,” pp. 1985–1988, 2012.
- [22] C. Chan and E. W. M. Yu, “An Abnormal Sound Detection and Classification System for Surveillance Applications,” pp. 1851–1855, 2010.
- [23] W. Liao and Y. Lin, “Classification of Non-Speech Human Sounds : Feature Selection and Snoring Sound Analysis,” no. October, pp. 2695–2700, 2009.
- [24] A. Fleury, N. Noury, M. Vacher, H. Glasson, and J.-F. Seri, “Sound and speech detection and classification in a Health Smart Home,” 2008 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 4644–4647, 2008.
- [25] C. Zhang and J. H. L. Hansen, “Analysis and Classification of Speech Mode : Whispered through Shouted,” pp. 1–4, 2007.
- [26] M. Davy and S. Rossignol, “Improved One-Class Svm Classifier For Sounds,” pp. 117–122, 2007.
- [27] P. C. Schön, B. Puppe, and G. Manteuffel, “Automated recording of stress vocalisations as a tool to document impaired welfare in pigs,” *Anim. Welf.*, vol. 13, no. 2, pp. 105–110, 2004.
- [28] M. Vacher and D. Istrate, “Sound detection and classification through transient models using wavelet coefficient trees,” pp. 1171–1174, 2004.
- [29] E. R. Siebert and L. A. Parr, “A Structural and Contextual Analysis of Chimpanzee Screams,” vol. 109, pp. 104–109, 2003.
- [30] N. E. O. Connor, J. Kuklyte, and P. Kelly, “Anti-social Behavior Detection in Audio-Visual Surveillance Systems,” no. June, pp. 40–41, 2009.
- [31] R. A. Breguet, “Automatic Sound Detection And Recognition For Noisy Environment,” 2000.
- [32] D. Lyon, *Surveillance studies : an overview*. Polity, 2007.
- [33] S. Lefèvre and N. Vincent, “A two level strategy for audio segmentation,” *Digit. Signal Process. A Rev. J.*, vol. 21, no. 2, pp. 270–277, Mar. 2011.
- [34] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
- [35] Lie Lu, Hong-Jiang Zhang, and Hao Jiang, “Content analysis for audio classification and segmentation,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.
- [36] N. Bhatia and Vandana, “Survey of Nearest Neighbor Techniques,” Jul. 2010.
- [37] I. Rojek and M. Jagodziński, “Hybrid Artificial Intelligence System in Constraint Based Scheduling of Integrated Manufacturing ERP Systems,” Springer, Berlin, Heidelberg, 2012, pp. 229–240.
- [38] D. Turnbull and C. Elkan, “Fast recognition of musical genres using RBF networks,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 580–584, Apr. 2005.
- [39] C. W. Tao, “A reduction approach for fuzzy rule bases of fuzzy controllers,” *IEEE Trans. Syst. Man Cybern. Part B*, vol. 32, no. 5, pp. 668–675, Oct. 2002.
- [40] A. Temko, D. Macho, and C. Nadeu, “Fuzzy integral based information fusion for classification of highly confusable non-speech sounds,” *Pattern Recognit.*, vol. 41, no. 5, pp. 1814–1823, May 2008.
- [41] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, “Violence Content Classification Using Audio Features,” Springer, Berlin, Heidelberg, 2006, pp. 502–507.
- [42] P. Prodanov and A. Drygajlo, “Bayesian networks based multi-modality fusion for error handling in human-robot dialogues under noisy conditions,” *Speech Commun.*, vol. 45, no. 3, pp. 231–248, Mar. 2005.
- [43] K. Daoudi, D. Fohr, and C. Antoine, “Dynamic Bayesian networks for multi-band automatic speech recognition,” *Comput. Speech Lang.*, vol. 17, no. 2–3, pp. 263–285, Apr. 2003.
- [44] R. A. Fisher, “The Use Of Multiple Measurements In Taxonomic Problems,” *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, Sep. 1936.
- [45] V. N. Vapnik, “Statistical learning theory New York,” 1998.
- [46] S. Sharma and R. L. Yadav, “Comparative Study of Kmeans and Robust Clustering,” *Int. J. Adv. Comput. Res.*, vol. 3, no. 12, 2013.
- [47] T. Andreassen, A. Surlykke, and J. Hallam, “Semi-automatic long-term acoustic surveying: A case study with bats,” *Ecol. Inform.*, vol. 21, pp. 13–24, May 2014.
- [48] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. Wiley-Interscience, 1990.
- [49] B. Larsen and C. Aone, “Fast and effective text mining using linear-time document clustering,” in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, 1999, pp. 16–22.
- [50] L. E. Baum and T. Petrie, “Statistical Inference for Probabilistic Functions of Finite State Markov Chains,” *The Annals of Mathematical Statistics*, vol. 37. Institute of Mathematical Statistics, pp. 1554–1563.
- [51] J. Principe, N. Euliano, and W. Lefebvre, *Neural and adaptive systems: fundamentals through simulations*, Vol.672. New York: Wiley, 2000.

Performance Improvement of Web Proxy Cache Replacement using Intelligent Greedy-Dual Approaches

Waleed Ali

Department of Information Technology
Faculty of Computing and Information Technology, King Abdulaziz University
Rabigh, Kingdom of Saudi Arabia

Abstract—This paper reports on how intelligent Greedy-Dual approaches based on supervised machine learning were used to improve the web proxy caching performance. The proposed intelligent Greedy-Dual approaches predict the significant web objects' demand for web proxy caching using Naïve Bayes (NB), decision tree (C4.5), or support vector machine (SVM) classifiers. Accordingly, the proposed intelligent Greedy-Dual approaches effectively make the cache replacement decision based on the trained classifiers. The trace-driven simulation results indicated that in terms of byte hit ratio and/or hit ratio, the performance of each of the conventional Greedy-Dual-Size-Frequency (GDSF) and Greedy-Dual-Size (GDS) was noticeably enhanced by applying the proposed Greedy-Dual approaches on five real datasets.

Keywords—Cache replacement; Greedy-Dual approaches; machine learning; proxy

I. INTRODUCTION

Internet performance can be improved by several approaches, any one of which may not always be the best method, due to practical issues such as network infrastructure, environment, and cost of hardware [1]. The second and the most popular approach is a web caching technique [1], [2], which decreases the network load by providing the requested web content from local storage. In a similar manner to caching in the cache memory to enhance CPU performance, web caching stores some web objects in anticipation of future requests, to enhance Web-based systems.

Basically, the implementation of web caching is done in three levels: client machine, proxy server and/or origin server. However, it is considered that the most significant caching approach is web proxy caching [2]-[7] which is used to save the networks' bandwidth, reduce Internet network traffic and decrease user-perceived latency.

In some situations, the proxy cache buffer is full of the stored web objects and a cache replacement policy is executed to provide enough space for the new incoming objects. The proxy cache replacement policy is responsible for removing unwanted web objects which may cause proxy cache pollution and poor performance.

Greedy-Dual-Size-Frequency (GDSF) and Greedy-Dual-Size (GDS) are two of the most commonly used web pages caching strategies, which are applied at proxy server. In GDS

and GDSF, the replacement cache decision is made based on mathematical equations combining a few important features of the object. Higher priority is given by GDS and GDSF to small web objects compared with large objects. Thus, the hit ratio is maximized, but at the expense of the byte hit ratio. Since web users' interests change depending on rapid changes in a web environment, smart and adaptive approaches are required to contribute to the web caching and replacement decisions.

II. SUMMARY OF CONTRIBUTIONS

Least-Frequently-Used-Dynamic-Aging (LFU-DA) and Least-Recently-Used (LRU) were enhanced using supervised machine learning in previous works [7], [8], respectively. However, the hit ratio measure achieved by the intelligent LRU and LFU-DA approaches were not good enough compared to GDS and GDSF because neither the size nor the retrieving time of web pages was considered in these approaches. In this paper it is shown how intelligent machine learning classifiers are effectively utilized in the GDS and GDSF in order to obtain optimal and intelligent Greedy-Dual approaches that can perform better in terms of both bytes hit ratio and hit ratio.

GDS is combined with intelligent machine learning classifiers to produce novel smart GDS caching methods (such as SVM-GDS, C4.5-GDS, and NB-GDS) with better performance. In the proposed intelligent GDS caching approaches, the frequency factor in the conventional GDS policy is replaced with the probability (computed by either the trained C4.5, SVM or NB classifier) of re-accessing the object soon.

In addition, C4.5, SVM or NB is incorporated with GDSF to improve the low byte hit ratio. The subsequent proposed replacement approaches are called C4.5-GDSF, SVM-GDSF and NB-GDSF. In the proposed intelligent GDSF approaches, the value of the object class (either one or zero) predicted by the trained classifier is added in the conventional GDSF in order to assign a higher priority to the web objects that are likely to be revisited soon.

The relative performances of the proposed intelligent Greedy-Dual approaches are then comprehensively discussed and compared with the most common and more relevant intelligent cache replacement methods.

The remainder of this paper is structured as follows. Section III describes the background of web proxy replacement and caching. Supervised machine learning is also presented briefly in subsection B while the current intelligent web cache replacement techniques are summarized in subsection C. Section IV presents the methodology of the proposed intelligent Greedy-Dual algorithms. The proposed approaches are evaluated and compared with other conventional and intelligent cache replacement techniques in Section V. Finally, Section VI concludes the work proposed in this study and suggests future work arising from this paper.

III. BACKGROUND AND RELATED WORK

A. Web Proxy Cache Replacement

The web proxy caching is a useful technique that plays an essential role in improving the performance of Web-based systems in terms of minimizing the utilization of network bandwidth, decreasing user-perceived delays and reducing loads on the original servers.

Three popular aspects have high impact on web proxy caching, which are cache consistency, cache pre-fetching, and cache replacement [1], [3], [4]. However, the powerful cache replacement method is essential and can make the greatest contribution in enhancing the caching performance [5]-[10].

When the proxy cache becomes full of web objects, a replacement strategy is basically used to manipulate the contents of the cache to provide sufficient space for incoming objects. The primary objective of the ideal cache replacement policy is to eliminate the undesired objects, to provide the best utilization of the proxy cache. Hence, cache hit rates can be improved, and loads on the server can be reduced.

A Greedy-Dual-Size (GDS) policy is suggested by [11] to lessen the cache pollution issues faced by the SIZE policy. In addition to the size factor, the cost of retrieving a web object from the server and the aging factor are combined with the key value assigned by GDS for each object available in the proxy cache. As the proxy cache is fully occupied, the web object that has the lowest key value is removed to provide enough place to the new demanded objects. The GDS policy uses (1) to compute $K(g)$, which represents the caching priority of object g visited by a web user.

$$K(g) = L + \frac{C(g)}{S(g)} \quad (1)$$

Where $S(g)$ is the size of g ; $C(g)$ is the fetching cost of g from its origin server; and L is an aging factor, which has the zero as the initial value and is then adjusted to the caching priority of the last replaced object.

When object g is requested again, $K(g)$ is modified based on the updated L value. Hence, the objects visited recently have larger caching priority values. The GDS policy obtains a much better hit ratio compared with other conventional replacement methods. However, the GDS approach still suffers from a low byte hit ratio [11]. Therefore, [12] suggested

an improvement on GDS by integrating the visit frequency $F(g)$ into the replacement decision, to produce Greedy-Dual-Size-Frequency (GDSF), as shown in (2). GDSF accomplishes a higher hit ratio compared to other cache replacement methods. However, although GDSF obtains a higher byte hit ratio than GDS, GDSF still performs minimal byte hit ratio compared to the other conventional replacement methods [12].

$$K(g) = L + F(g) * \frac{C(g)}{S(g)} \quad (2)$$

B. Supervised Machine Learning

The supervised learning algorithm works on the training dataset to generate a classifier that has the ability to predicting the correct class for the known dataset (testing dataset). This section concentrates on three popular machine learning algorithms: decision tree (C4.5), support vector machine (SVM) and Naïve Bayes classifier (NB), which have been successfully applied in many applications [13]-[17].

In the decision tree, a feature in the training instance is represented by a node, while each tree branch has a value, which can be predicted by that node. The C4.5 developed by [17] is the most commonly used algorithm to generate a decision tree for classification purposes. The C4.5 is constructed based on a top-down recursive approach to generate the decision tree. All of the training instances are initially at the tree root. The C4.5 then uses an impurity function in order to split the training instances recursively. The partitioning process is then repeated until all instances for a given node belong to the same class.

A support vector machine, which is a discriminative model, aims to achieve an optimal hyperplane which categorizes new instances by generating the maximal likely distance between the separating hyperplane and the instances in order to decrease the upper bound on the predictable generalization error. In the SVM training, support vectors closer to the separating hyperplane are obtained from the dataset to represent the most valuable instances used for classification. In addition to linear classification, SVMs can be used to solve other non-linear classification problems by selecting the appropriate kernel function to convert the instances into high-dimensional spaces.

One of the simplest Bayesian networks is the Naive Bayes network (NB), which is represented as a directed acyclic graph in which the class label is represented by the single parent and the features are represented by some children. NB supposes that no correlation exists between the features and that, given the class label, all the features are conditionally independent. The conditional probabilities $\Pr(A_i = a_i | C = c_j)$ and the prior probabilities $\Pr(C = c_j)$ are computed in the training phase. Formula (3) is then used in order to predict the class of a test example.

$$c = \arg \max_{c_j} \Pr(C = c_j) \prod_{i=1}^{|A|} \Pr(A_i = a_i | C = c_j) \quad (3)$$

TABLE I. SUMMARY OF THE EXISTING INTELLIGENT WEB CACHE REPLACEMENT TECHNIQUES

Machine Learning Used	Existing Works	Based on	Cache Location	Data Used for Evaluation
SVM	SVM-DA [7]	LFU-DA	Proxy	The IRCache network's proxy logs files
	SVM-LRU [8]	LRU	Proxy	The IRCache network's proxy logs files
Decision Tree	C4.5-DA [7]	LFU-DA	Proxy	The IRCache network's proxy logs files
	C4.5-LRU [8]	LRU	Proxy	The IRCache network's proxy logs files
	J48-C-LRU [22]	LRU	Proxy	The IRCache network's proxy logs files
	CART, MARS, RF and TN in web caching [23]	LRU	Client and Sever	- Cunha of Boston University 's Web client traces - E-Learning @UTM Web server
Naïve Bayes Classifier	NB-DA [7]	LFU-DA	Proxy	The IRCache network's proxy logs files
	NB-LRU [8]	LRU	Proxy	The IRCache network's proxy logs files
ANN	NNPCR [24] and NNPCR-2 [6]	LFU-DA	Proxy	The IRCache network's proxy logs files
	BP and PSO in Web caching [25]	None	Server	Cunha of Boston University 's Web client traces
	LRU-C [9]	LRU	Server	Finnish University and Research Network access's logs file
ANFIS	ICWCS [26]	LRU	Client	Cunha of Boston University 's Web client traces
Logistic Regression	LRU-C and LRU-M [27]	LRU	Proxy	The IRCache network's proxy logs files just for one day
	Logistic regression in an adaptive web cache [28]	None	Server	Server logs files's Internet Traffic Archive

C. Related Works on Intelligent Web Cache Replacement Techniques

Several intelligent methods have been explored as alternative solutions to enhance the performance of traditional approaches of proxy cache replacement. The intelligent proxy cache replacement methods have been developed by using supervised machine learning techniques (see Table I), fuzzy systems [18], or evolutionary algorithms [19]-[21]. The existing intelligent web cache replacement techniques based on the supervised machine learning are considered as the most commonly used, effective and adaptive approaches, as summarized in Table I.

By examining the existing works cited in Table I, it can be concluded that two intelligent replacement paradigms are dominant in the existing intelligent web cache replacement techniques. A supervised machine learning technique is utilized independently in the proxy cache replacement or incorporated with one of the conventional replacement policies such as LFU-DA or LRU. The object size and cost are not considered in the replacement decision with these paradigms.

Unlike the previous works, the proposed intelligent Greedy-Dual approaches can remarkably enhance the byte hit ratio of the conventional GDSF and GDS. Besides, they utilize the advantages of GDS and GDSF in terms of high hit ratio. In other words, intelligent machine learning classifiers are effectively utilized into the GDS and GDSF in order to obtain optimal intelligent Greedy-Dual approaches that can achieve good performance in both the hit ratio and the byte hit ratio.

IV. METHODOLOGY

A methodology for enhancing web proxy cache replacement using intelligent Greedy-Dual approaches is explained in this section. The methodology involves two phases: training of supervised machine learning classifiers, and then integrating the trained classifiers into the web proxy cache replacement.

A. Training of Supervised Machine Learning Classifiers

In order to effectively predict the desired web object, C4.5, SVM and NB classifiers are trained with training data prepared based on users' requests recorded in the web proxy logs file. Some features of the training dataset are extracted from the web proxy logs file immediately, while other features are prepared using equations, as shown in Table II. The target output for each request is also prepared from the proxy logs file, based on the forward-looking sliding window (SWL) as shown in (4).

$$y = \begin{cases} 1, & \text{if the object would be revisited within forward SWL} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

As can be observed, the input features are based on the past information of objects requests within the backward-looking sliding window to expect whether these objects would be revisited soon or not within the forward-looking sliding window.

TABLE II. THE FEATURES PREPARATION OF TRAINING DATASET

Feature Name	Description	How to Prepare
SWL-based Recency	Recency of visiting a object withing backward-looking sliding window	$x_1 = \begin{cases} \text{Max}(SWL, \Delta T) & , \text{if object } g \\ & \text{was requested before} \\ SWL & , \text{otherwise} \end{cases}$ where ΔT is the time in seconds since object g was last request , and SWL is sliding window length.
Frequency	Visits Frequency of a object	Number of requests for a web object in proxy logs file
SWL-based Frequency	Visits Frequency of a object within backward-looking sliding window	$x_{3,i} = \begin{cases} x_{3,i-1} + 1 & , \text{if } \Delta T \leq SWL \\ 1 & , \text{otherwise} \end{cases}$
Retrieval time	fetching time of a object in milliseconds	extracted from elapsed time field of log entry in the proxy logs file
Size	Size of object in bytes	extracted from size field of log entry in the proxy logs file
Type	Type of web object	1 for HTML, 2 for image, 3 for audio, 4 for video, 5 for application and zero for others.

When the proxy dataset is preprocessed well, C4.5, SVM and NB can be trained using the prepared dataset for web object classification. The training phase aims to train C4.5, SVM and NB classifiers to predict the web object class requested by the user, either as objects to be revisited soon or not. Consequently, the classification information is utilized with the cache replacement decision to enhance the web proxy caching performance.

B. Proposed Intelligent Greedy-Dual Approaches

As NB, C4.5 and SVM are correctly trained to classify proxy cache contents, as discussed earlier; a web proxy cache replacement strategy can utilize NB, C4.5 or SVM classifiers for managing the contents of the web proxy cache. As shown in Fig. 1, when a web user visits object g , the cache manager searches for object g in the proxy cache. Whether a cache hit or miss has occurred, intelligent Greedy-Dual approaches are used to compute or update the caching priority, $K(g)$, of g . The desired features of g , as shown in Table II, are collected and utilized as inputs for the classification algorithm that can classify object g as an object that would be revisited again or not. Thus, the classification decision is incorporated into the GDS or GDSF cache replacement approach for updating the priority of g . Then, g is reordered and located depending on the new priority of g in the cache list. Consequently, the proposed intelligent GDS and GDSF can identify and remove the unwanted web objects with the lowest priority for replacement.

In the proposed intelligent GDS approaches, classification information produced by C4.5, SVM or NB classifier is combined with the conventional GDS to enhance the byte hit ratio. The suggested intelligent GDS approaches are so-called NB-GDS, C4.5-GDS and SVM-GDS. In the proposed intelligent GDS approaches, a NB, C4.5 or SVM classifier is used to compute the probability, $Pr(g)$, of revisiting object g in the near future. Each time a user visits an object g , the accumulated $Pr(g)$, i.e., $W(g) = \sum Pr(g)$, is combined with the caching priority $K(g)$ using (5).

$$K(g) = L + W(g) * \frac{C(g)}{S(g)} \quad (5)$$

In addition to the intelligent GDS, the traditional GDSF is extended based on a NB, C4.5 or SVM classifier to enhance the low byte hit ratio. Therefore, the proposed NB-GDSF, C4.5-GDSF and SVM-GDSF are produced as alternative approaches to the traditional GDSF web proxy cache replacement method.

In the proposed intelligent GDSF approaches, the trained NB, C4.5 or SVM classifier is applied for the prediction of the web objects' class (one or zero) requested by the web user. The class label is then included as an additional weight into GDSF to provide higher priority to the preferred objects, which will be revisited sometime in future even if the preferred objects are large. When a web user visits g , the intelligent GDSF uses (6) to assign the caching priority, $K(g)$, of object g . Hence, based on its priority, g is relocated in the proxy cache.

$$K(g) = L + F(g) * \frac{C(g)}{S(g)} + W(g) \quad (6)$$

Where $W(g)$ is either one or zero, which represents the class of object g obtained using the NB, C4.5 or SVM classifier.

The rationale behind the proposed intelligent Greedy-Dual approaches is explained as follows. The conventional GDS and GDSF give greater priority to small web objects, which are removed first from the proxy cache. Thus, the hit ratio is maximized by the conventional GDS and GDSF but at the expense of the byte hit ratio. Instead of that, the suggested intelligent Greedy-Dual approaches can predict either the class value or probability of the preferred objects, which would be re-accessed soon using SVM, NB and C4.5 classifiers. Accordingly, the class information is successfully integrated with the storing priority of the web object. In other words, the priority values of those preferred objects can be enhanced using a SVM, NB or C4.5 classifier, regardless of their size and visits frequency. Thus, the proposed intelligent Greedy-Dual approaches can outstandingly enhance the byte hit ratio of the conventional GDS and GDSF. In addition, the superior hit ratio of the conventional GDS and GDSF can be maintained in the intelligent Greedy-Dual approaches.

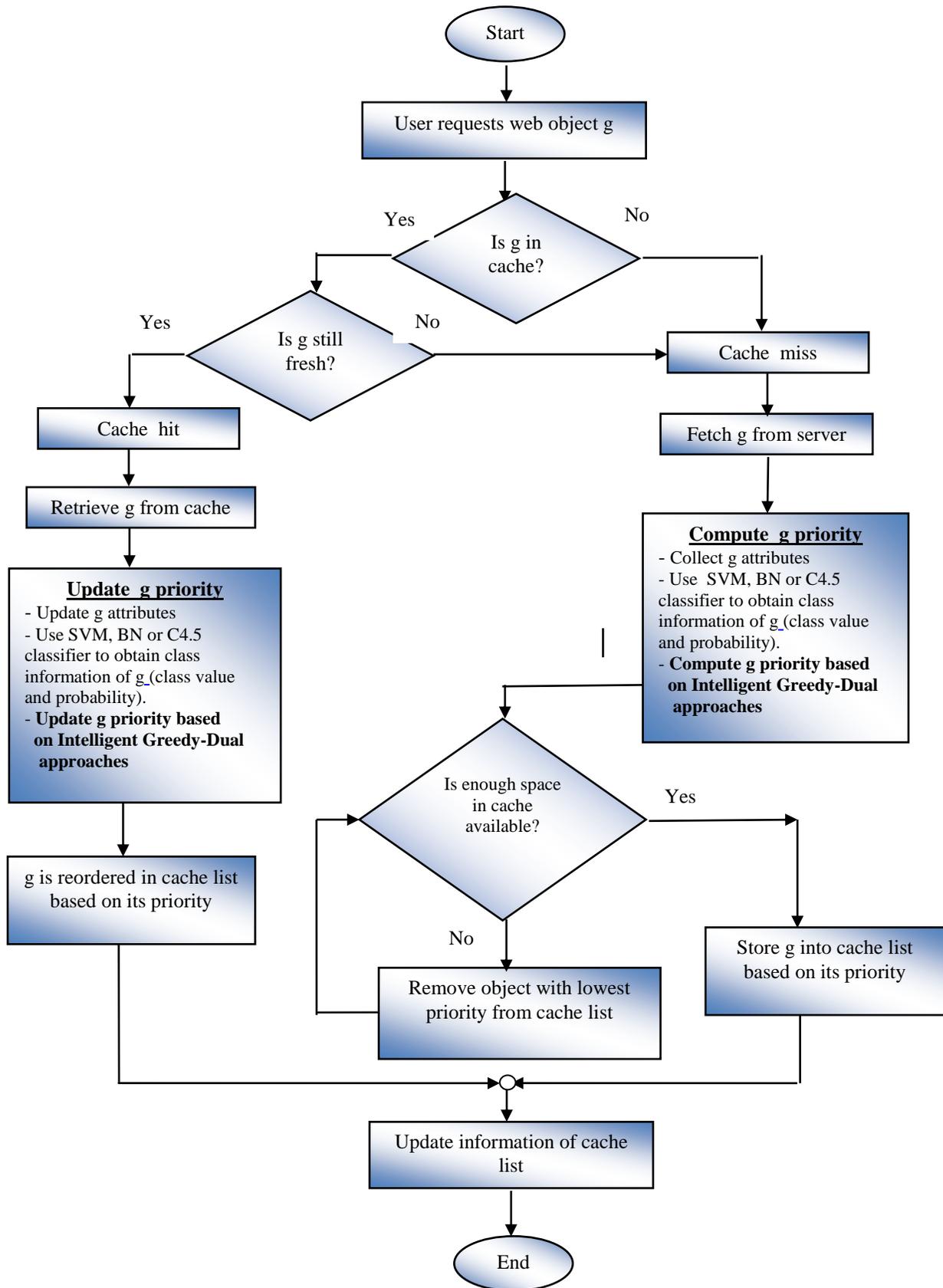


Fig. 1. A methodology for enhancing web proxy cache replacement using intelligent greedy-dual approaches.

V. RESULTS AND DISCUSSION

A. Data Collection

The proxy log files used in this study were obtained from five proxy servers (BO2, NY, UC, SV and SD) from the IRTCache network [29] that are located in the United States over a period of fifteen days. C4.5, SVM and NB classifiers were trained based on the data collected in the first day, while the remaining data of the two weeks were used to evaluate the suggested intelligent Greedy-Dual method against existing works.

B. Improvement Ratio of Hit and Byte Hit Ratio

In this study, a WebTraff [30] simulator was adjusted to simulate and evaluate the effectiveness of the performance of the proposed intelligent Greedy-Dual approaches against various existing web cache replacement policies.

The most popular measures used to verify and evaluate the performance of proxy cache replacement are hit ratio (HR) and byte hit ratio (BHR), which are related with the number of user's requests and bytes served by the proxy cache instead of the original server. Due to space limitations, (7) is used to calculate the average improvement ratios (*IRs*) of conventional method (*CM*) in terms of the HR and BHR obtained by the proposed method (*PM*), i.e., the intelligent GDS and GDSF against conventional GDS and GDSF.

$$IR = \frac{(PM - CM)}{CM} \times 100 (\%) \quad (7)$$

For the five datasets, Table III summarizes the average *IRs* performed by intelligent GDS approaches over conventional GDS for each particular cache size. The averages *IRs* were significantly influenced when the proxy cache size increased. More particularly, the impact of the performance of a replacement policy for the small cache was noticed clearly, since the replacement process occurred frequently.

For HR, the results show that SVM-GDS, NB-GDS and C4.5-GDS improved the HR of GDS with average *IRs* by up to 17.42%, 22.45% and 18.79% respectively, as shown in Table III. For the average *IRs* of the BHR, the BHR of the GDS was significantly enhanced by SVM-GDS, NB-GDS and C4.5-GDS, by up to 57.61%, 229.14% and 85.65%, respectively. This was mainly due to the capability of intelligent GDS approaches to intelligently remove the correct objects from the proxy cache. By contrast, the low BHR of the conventional GDS was expected, due to the GDS's weighting toward smaller objects, even if the smaller objects are not preferred.

From Table III, it can also be seen that the HR of C4.5-GDS was almost the same as the HR of SVM-GDS, but slightly lower than that of NB-GDS. In terms of the BHR, NB-GDS accomplished the best BHR, while SVM-GDS attained the worst BHR compared to the BHRs of NB-GDS and C4.5-GDS. This was due to the fact that NB-GDS gave more

accurate probabilities or scores to the preferred objects, either small or large objects. This contributed greatly to obtaining a good HR and a much better BHR from NB-GDS than from the others.

The average *IRs* achieved by the intelligent GDSF methods are also presented in Table III. SVM-GDSF, NB-GDSF and C4.5-GDSF accomplished good HRs but these were slightly inferior to the HR of the conventional GDSF. In the worst case, SVM-GDSF, NB-GDSF and C4.5-GDSF lost 7.29%, 9.43% and 7.4% respectively from the HR of GDSF. However, the BHR of GDSF was significantly enhanced by SVM-GDSF, GDSF-NB and C4.5-GDSF and increased by 407.49%, 380.55%, and 308.08%, respectively. This enhancement was obtained because the GDSF tends to cache many of the small objects in the proxy cache to increase the HR, but at the expense of BHR.

Table III shows also that C4.5-GDSF and SVM-GDSF achieved slightly higher HRs than the HR of NB-GDSF, while NB-GDSF and SVM-GDSF achieved better BHRs compared to the BHR of C4.5-GDSF. This meant that the best balance between the HR and BHR was achieved by SVM-GDSF.

C. Overall Comparison and Discussion

As shown in the previous section, the proposed NB-GDS and SVM-GDSF achieved a more competitive HR and better BHR. Thus, NB-GDS and SVM-GDSF were selected to be used in the overall comparison. The proposed NB-GDS and SVM-GDSF approaches were compared with the most common cache replacement methods used in squid software such as LRU, GD, GDSF and LFU-DA [24], [6]. In addition, NB-GDS and SVM-GDSF were compared with other existing intelligent proxy cache replacement methods, such as NNPCR-2 [6], SVM-LRU [8], and SVM-DA [7].

In terms of the HR, Fig. 2 clearly indicates that SVM-LRU, NB-GDS and SVM-DA improved the performance of LRU, GDS and LFU-DA, respectively on the five proxy datasets. Conversely, the HR of SVM-GDSF was comparatively or somewhat worse than the HR of GDSF. Fig. 2 also demonstrates that the HRs of NB-GDS, SVM-GDSF and SVM-DA were much better than the HR of NNPCR-2, while the HR of SVM-LRU was slightly better than that of NNPCR-2 for most of the proxy datasets. From Fig. 2, it can be concluded that the best HR was achieved by NB-GDS, while the worst HR was given by LRU on all datasets.

In terms of BHR, Fig. 3 demonstrates that, for all proxy datasets, the BHR obtained by GDS and GDSF was much lower than that achieved by LFU-DA, LRU and NNPCR-2. This was expected, since LFU-DA, LRU and NNPCR-2 policies removed objects regardless of their sizes. Furthermore, the BHRs of SVM-DA and SVM-LRU were better than those of LFU-DA, LRU and NNPCR-2 in all proxy datasets with different cache sizes.

It can also be noticed from Fig. 2 and 3 that although GDS and GDSF had a better a performance for the HR obtained compared to the others, it can clearly be seen that the BHRs of GDS and GDSF were the worst among all the methods. This is because GDS and GDSF prefer to cache the small and recent objects.

TABLE III. THE AVERAGE IRs ACHIEVED BY INTELLIGENT GDS AND GDSF OVER GDS AND GDSF

Cache Size (MB)	Average IR of HR and BHR Over GDS and GDSF (%)											
	SVM-GDS		SVM-GDSF		NB-GDS		NB-GDSF		C4.5-GDS		C4.5-GDSF	
	HR	BHR	HR	BHR	HR	BHR	HR	BHR	HR	BHR	HR	BHR
1	17.42	26.23	-1.59	16.42	22.45	46.47	-0.86	16.85	18.79	33.96	-1.03	16.34
2	15.87	26.37	-3.13	27.31	18.96	51.38	-4.55	105.22	17.05	35.21	-2.46	20.43
4	10.56	23.95	-3.88	45.61	15.21	96.64	-10.81	255.28	14.52	28.70	-6.07	32.85
8	13.15	41.09	-5.37	155.77	15.25	229.14	-6.97	278.46	13.97	56	-3.27	125.2
16	11.05	48.47	-6.01	123.57	12.92	179.98	-6.47	147.21	11.6	85.65	-4.35	91.98
32	10.09	54.44	-6.50	94.15	11.31	125.3	-7.76	96.59	10.45	81.01	-5.99	76.7
64	9.74	49.08	-6.39	407.49	10.7	202.07	-7.79	380.55	9.95	54.82	-6.18	308.08
128	7.08	57.61	-7.28	125.95	8.77	105.77	-9.43	96.45	7.29	73.1	-7.4	93.04
256	5.11	50.59	-6.96	84.55	7.65	88.36	-8.79	65.61	5.32	59.39	-6.22	54.34
512	2.96	44.78	-5.32	61.66	5.58	75	-8.18	38.09	3.09	59.61	-4.88	34.42
1024	1.95	35.22	-4.13	40.8	5.32	71.87	-6.44	25.29	2.23	51.34	-2.94	19.38
2048	0.64	34.71	-2.50	25.59	4.47	55.11	-5.11	17.18	0.87	54.97	-2.63	8.15
4096	0.28	28.75	-2.49	9.16	4.41	30.82	-4.08	10.23	0.40	29.89	-1.61	4.81
8192	0.15	8.17	-0.37	2.54	3.86	8.66	-2.06	2.40	0.20	8.42	-0.69	1.95
16384	0.03	1.27	-0.05	1.16	3.89	1.56	-0.25	0.24	0.05	1.58	-0.20	0.12
32768	0	0.56	0	0.02	3.87	0.96	0	0.33	0	0.94	-0.02	0.28

Fig. 2 and 3 show that both SVM-GDSF and NB-GDS were significantly improved in terms of BHRs achieved over GDSF and GDS, respectively. It can be concluded that the proposed NB-GDS and SVM-GDSF achieved outstanding HRs and competitive BHRs for most of the proxy datasets.

VI. CONCLUSION AND FUTURE WORK

In this paper, intelligent Greedy-Dual approaches have been suggested to obtain optimal web proxy cache replacement approaches that can achieve good performance in both HR and BHR. To improve the lower byte hit ratios of the conventional GDS and GDSF policies, intelligent machine learning classifiers were combined with these policies to produce novel intelligent GDS and GDSF caching approaches with better performance. The trace-driven simulation results depicted that the intelligent Greedy-Dual approaches noticeably enhanced the performance of the traditional GDS in terms of byte hit ratio and hit ratio. The averages of the IRs of the BHRs obtained by SVM-GDS, NB-GDS and C4.5-GDS over GDS increased by 57.61%, 229.14% and 85.65%, respectively, while the average IRs of the HR increased by 17.42%, 22.45% and 18.79%, respectively. Moreover, the intelligent GDSF approaches significantly improved the performance in terms of

the byte hit ratio of GDSF. The average IRs of the BHRs of SVM-GDSF, NB-GDSF, and C4.5-GDSF were many times greater than the BHRs of GDSF, and increased by 407.49%, 380.55%, and 308.08%, respectively. When the proposed intelligent Greedy-Dual approaches were compared with conventional and other intelligent replacement approaches, it was observed that the proposed NB-GDS achieved the best HR. Furthermore, BHRs of SVM-GDSF and NB-GDS were competitive with the BHRs of LRU and LFU-DA for most proxy datasets.

The proposed intelligent Greedy-Dual approaches can be implemented in real environments such as organizations or universities. For example, the proposed approaches can be implemented on proxy servers of departments, faculties and campus, to reduce the response time and save the network bandwidth of server. The proposed approaches do not consider multiple caching proxies, which cooperate and share their caches. In addition, regular retraining of classifiers is expected to improve the adaptability and efficiency of the proposed intelligent caching approaches. Eventually, instead of stand-alone web caching, intelligent Greedy-Dual approaches can be effectively integrated with a prefetching policy in order to improve the web performance.

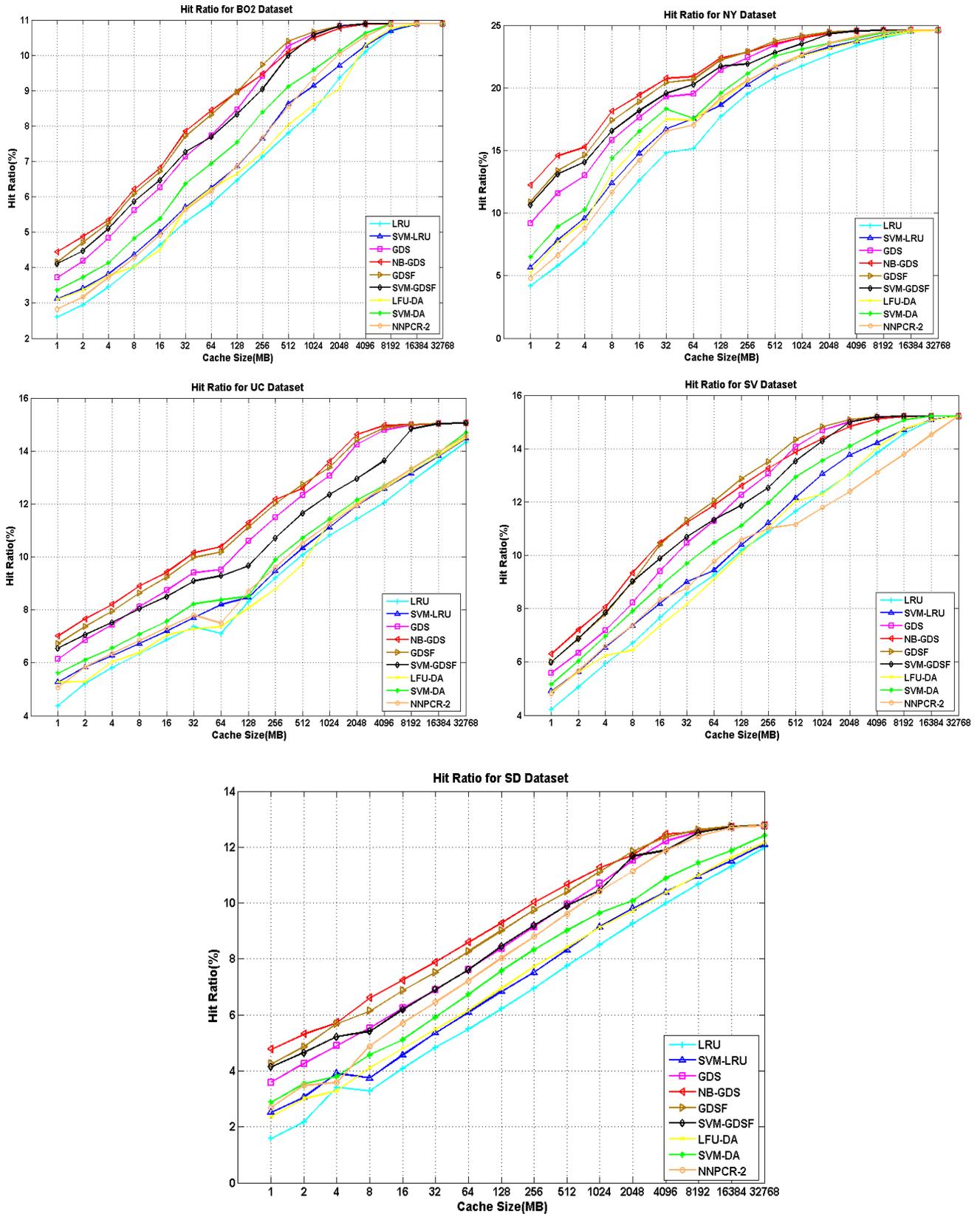


Fig. 2. Comparison of hit ratio between the conventional and intelligent web proxy caching approaches.

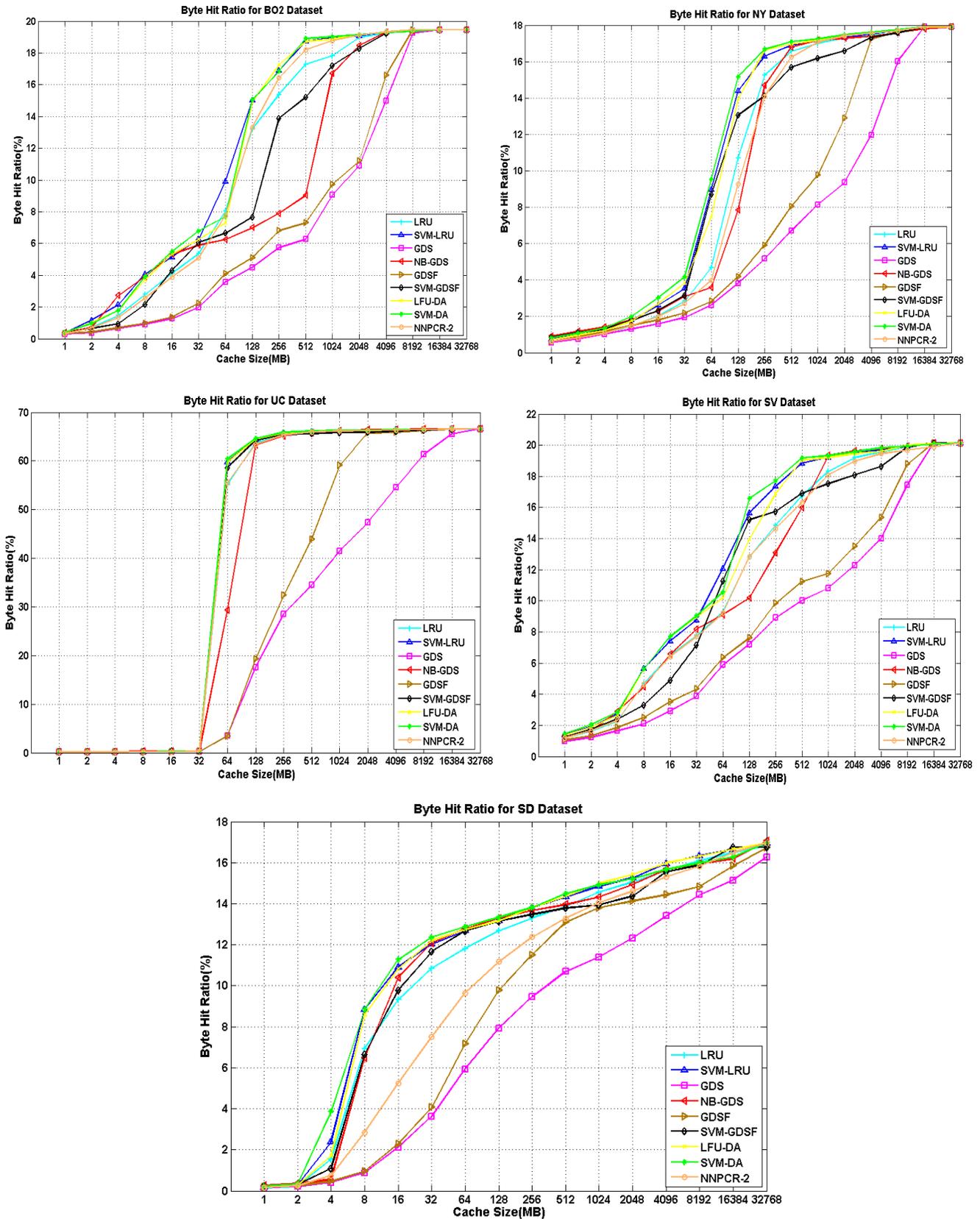


Fig. 3. Comparison of byte hit ratio between the conventional and intelligent web proxy caching.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant No. (830-751-D1435). The authors, therefore, gratefully acknowledge the DSR technical and financial support.

REFERENCES

- [1] U. Acharjee, "Personalized and Artificial Intelligence Web Caching and Prefetching," MSc, University of Ottawa, Canada, 2006.
- [2] C. Kumar, and J. B Norris, "A new approach for a proxy-level web caching mechanism," *Decision Support Systems*, vol. 46(1), pp. 52-60, 2008.
- [3] C. Kumar, "Performance evaluation for implementations of a network of proxy caches," *Decision Support Systems*, vol. 46(2), pp. 492-500, 2009.
- [4] C. C. Kaya, G. Zhang, Y. Tan, and V. S. Mookerjee, "An admission-control technique for delay reduction in proxy caching," *Decision Support Systems*, vol. 46(2), pp. 594-603, (2009).
- [5] H. ElAarag, "A quantitative study of web cache replacement strategies using simulation," In *Web Proxy Cache Replacement Strategies*, Springer, London, pp. 17-60, 2013.
- [6] S. Romano, and H. ElAarag, "A neural network proxy cache replacement strategy and its implementation in the Squid proxy server," *Neural Computing & Applications*, vol. 20, pp. 59-78, 2011.
- [7] W. Ali, and S.M. Shamsuddin, "Intelligent Dynamic Aging Approaches in Web Proxy Cache Replacement," *Journal of Intelligent Learning Systems and Applications*, vol. 7(4), pp.117-127, 2015.
- [8] W. Ali, S. Sulaiman, and N. Ahmad, "Performance Improvement of Least-Recently-Used Policy in Web Proxy Cache Replacement Using Supervised Machine Learning," *International Journal of Advances in Soft Computing & Its Applications*, vol. 6(1), pp.1-38, 2014.
- [9] T. Koskela, J. Heikkonen, and K. Kaski, "Web cache optimization with nonlinear model using object features," *Computer Networks*, vol. 43, pp. 805-817, 2003.
- [10] T. Chen, "Obtaining the optimal cache document replacement policy for the caching system of an EC website," *European Journal of Operational Research*, vol. 181(2), pp. 828-841, 2007.
- [11] P. Cao, and S. Irani, "Cost-Aware WWW Proxy Caching Algorithms," In *Usenix symposium on internet technologies and systems*, Monterey, CA: pp. 193-206, 1997.
- [12] L. Cherkasova, "Improving WWW proxies performance with greedy-dual-size-frequency caching policy," *Hewlett-Packard Laboratories*, 1998.
- [13] R.C. Chen, and C.H. Hsieh, "Web page classification based on a support vector machine using a weighted vote schema," *Expert Systems with Applications*, vol. 31(2), pp. 427-435, 2006.
- [14] SH. Lu, DA. Chiang, HC. Keh, and HH. Huang, "Chinese text classification by the Naive Bayes Classifier and the associative classifier with multiple confidence threshold values," *Knowledge-Based Systems*, vol. 23(6), pp. 598-604, 2010.
- [15] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY, and Zhou ZH, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14(1), pp.1-37, 2008.
- [16] CJ. Huang, YW. Wang, TH . Huang, CF. Lin, CY. Li, HM. Chen, PC. Chen, and JJ. Liao, "Applications of machine learning techniques to a sensor-network-based prosthesis training system," *Applied Soft Computing*, vol. 11(3), pp. 3229-3237, 2011.
- [17] JR. Quinlan, "C4.5: Programming for machine learning.. Morgan Kauffmann, 1993.
- [18] MC. Calzarossa, and G. Valli, "A fuzzy Algorithm for web caching," *SIMULATION SERIES*, vol. 35(4), pp. 630-636, 2003.
- [19] K. Tirdad , F. Pakzad, and A. Abhari, "Cache replacement solutions by evolutionary computing technique," In *Proceedings of the 2009 Spring Simulation Multi conference*, San Diego, California: International Society for Computer Simulation. p. 128, 2009.
- [20] A. Vakali, "Evolutionary techniques for Web caching," *Distributed and Parallel Databases*, vol. 11(1), pp. 93-116, 2002.
- [21] C. Yan, L. Zeng-Zhi, and W. Zhi-Wen, "A GA-based cache replacement policy," In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, 2004.
- [22] W. Yasin, H. Ibrahim, NI. Udzir, and NA. Hamid, "Intelligent Cooperative Least Recently Used Web Caching Policy based on J48 Classifier," In *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services 2014 Dec 4* (pp. 262-269), ACM, 2014.
- [23] S. Sulaiman, S.M. Shamsuddin, and A. Abraham, "Intelligent web caching using machine learning methods," *Neural Network World*, vol. 21(5), pp. 429-452, 2011.
- [24] J. Cobb, and H. ElAarag, "Web proxy cache replacement scheme based on back-propagation neural network," *Journal of Systems and Software*, vol. 81, pp. 1539-1558, 2008.
- [25] S. Sulaiman, SM. Shamsuddin, F. Forkan, and A. Abraham, "Intelligent Web caching using neurocomputing and particle swarm optimization algorithm," In *2008 Second Asia International Conference on Modelling and Simulation (AICMS)*; 2008 May 13; Kuala Lumpur: IEEE. pp. 642-647, 2008.
- [26] A. A W, and S.M Shamsuddin, "Neuro-fuzzy system in partitioned client-side Web cache," *Expert Systems with Applications*, vol. 38, pp. 14715-14725, 2011.
- [27] G. Sajeev, and M. Sebastian, "A novel content classification scheme for web caches," *Evolving Systems*, vol. 2, pp. 101-118, 2011.
- [28] A.P Foong, H. Yu-Hen, and D.M Heisey, "Logistic regression in an adaptive Web cache," *Internet Computing*, vol. 3, pp. 27-36, 1999.
- [29] National Lab of Applied Network Research (NLANR) (2010) Sanitized access logs: Data collected between 21st August and 4th September , 2010, <http://www.ircache.net/>.
- [30] N. Markatchev, and C. Williamson, "Webtraff: A GUI for web proxy cache workload modeling and analysis," In *10th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems*; IEEE. pp. 356-363, 2002.

El Niño / La Niña Identification based on Takens Reconstruction Theory

Kohei Arai

Graduate School of Science and Engineering
Saga University
Saga City, Japan

Kaname Seto

Former Student, Graduate School of Science and
Engineering, Saga University
Saga City, Japan

Abstract—An identification method for earth observation data according to a chaotic behavior based on Takens reconstruction theory is proposed. The proposed method is examined by using the observed time series data of SST (Sea Surface Temperature) and the SOI (Southern Oscillation Index) data. The experimental results show that the time for the identification of the proposed method is not later than that of the existing method. Author confirmed that by using the definitions of the Japan Meteorological Agency and the use of Equations, I can identify El Niño / La Niña at an earlier time. In other words, we do not necessarily need a numerical value for 10 months by identifying the proposed method. I confirmed that the time required for the identification judgment of the proposed method is about one month. The proposed method is not based on extrapolation method with numerical model or governing equation, but based on interpolation method using only actual observation time series.

Keywords—Time series analysis; takens; sea surface temperature; SST; southern oscillation index; SOI; El Niño-southern oscillation; ENSO

I. INTRODUCTION

El Niño phenomenon / La Niña phenomenon¹ has been noticed not only around the equator but also as one of the things affecting the weather in the world numerical prediction and estimation of the phenomenon using numerical models and governing equations. Also, in Japan, the Meteorological Agency has made a judgment of the El Niño phenomenon / La Niña phenomenon. By the way, the definition of the El Niño phenomenon / La Niña phenomenon by the Meteorological Agency is defined as the El Niño surveillance area in the western coast of the eastern Pacific equatorial region from 4° North to 4° South and 150° West to 90° West.

It is assumed that the 5 month moving average value of the difference from the average sea surface temperature standard value (30 years average from 1961 to 1990) has reached 0.5° C or more (-0.5° C. or less) continuously for 6 months or more. In other words, data for minimum (2 + 6 + 2) months is necessary for judgment by the Meteorological Agency of the El Niño phenomenon / La Niña phenomenon. If the 5-month moving average value of the difference with the monthly average sea surface temperature standard value of the survey area becomes 0.5° C or more (-0.5° C or less) consecutively for less than 6 months, non-La Niña. It is judged as "Niña").

El Niño - pacific trade wind and extraordinary high water temperature in eastern equatorial region is discussed [1]. Also, a new wave of climate research El Niño and Southern Oscillation [2] is reported. On the other hand, global aspects of ENSO (El Niño / Southern Oscillation) [3] is discussed together with global structure of the El Niño / Southern Oscillation-Part I [4] and II [5]. Meanwhile, abrupt enhancement of convective, activity and low-level westerly burst during the onset phase of the 1986-87 El Niño [6] is discussed.

Anomalously short duration of the easterly wind phase of the QBO: Quasi-Biennial Oscillation² at 50 hPa in 1987 and its relationship [7] is discussed. Meanwhile, development of a twin cyclone and westerly, bursts during the initial phase of the 1986-87 El Niño [8] is well reported. Meantime, efforts to solve the El Niño phenomenon [9] is discussed. El Niño and prediction of extreme weather [10] is reported.

Statistical features of temperature and precipitation in the world during the El Niño La Niña phenomenon period [11] is reported. A research toward the prediction of the El Niño phenomenon [12] is also well reported. The current state of research on prediction of El Niño and related abnormal weather [13] is summarized in Central Research Institute of Electric Power Industry Research Report T, 98056, pp. 1 - 23, 1999. "Monitoring and prediction of global warming and El Niño phenomenon," is published by Meteorological Agency Climate Oceanic Weather Department Climate Information Division [14].

Use of The Tropical Ocean Global Atmosphere program (TOGA) TAO / TRITON buoy³ data⁴ for monitoring and prediction of El Niño phenomenon [15] is reported. Meanwhile, a study on the interaction between the atmosphere and the ocean using satellite observations and numerical models [16] is conducted. Simulation of typhoon and El Niño Southern Oscillation by high resolution atmosphere and ocean binding model [17] is well reported. Also, large scale meteorological phenomena in tropical zone and The Tropical Rainfall Measuring Mission TRMM⁵ observation - 1997-98 El Niño's termination and Madden-Julian vibration [18] is introduced.

² https://www.jstage.jst.go.jp/article/mripapers/48/1/48_1_1/_article

³ <https://www.pmel.noaa.gov/gtmba/>

⁴ <https://climatedataguide.ucar.edu/climate-data/tropical-moored-buoy-system-tao-triton-pirata-rama-toga>

⁵ <https://trmm.gsfc.nasa.gov/>

¹ <https://oceanservice.noaa.gov/facts/ninonina.html>

About the work of the El Niño monitoring center [19] is reported. On the other hand, El Niño surveillance center [20] is also reported. Meteorological Agency El Niño Observation Forecast Center: reported "El Niño events occurred," [21]. Meanwhile, about the updated "El Niño surveillance bulletin" [22] is reported.

El Niño Chaos is introduced in the Mathematical Science society [23]. Furthermore, numerical modeling for prediction El Niño's dynamic [24] is proposed. Meanwhile, detecting strange attractors in turbulence in Dynamical Systems and Turbulence [25] is discussed. The Takens embedding theorem⁶ is introduced [26].

El Niño phenomena analysis with Earth Observation Satellite data [27] is discussed. Also, El Niño and La Niña discrimination based on Takens reconstruction theory⁷ is proposed [28]. Prediction method of ENSO: El Niño Southern Oscillation⁸ event by means of wavelet based data compression with appropriate support length of base function [29] is also proposed.

The purpose of this research is to shorten the time required for judgment by the definition of the Japan Meteorological Agency. More generally, information on how long the state in which the absolute value of the 5-month moving average value of the difference from the reference value of the monthly average sea surface temperature in the surveillance area is α degree C. or more continues is obtained early Establishing a method.

In this paper, Takens reconstruction theorem possessing the feature that it is possible to extract its data characteristics only from target observation data having nonlinearity, and the feature that extrapolation problem can be transformed into interpolation problem A method for judging the La Niña phenomenon / La Niña phenomenon is proposed. Also, the effectiveness of the proposed method is validated in this paper.

The next section describes the proposed method together with typical conventional Newton-Raphson method⁹ for retrieving vertical profiles followed by experiments. Then conclusion with some discussions is followed by together with future research works.

II. PROPOSED METHOD

Consider extracting its dynamic characteristics from actual observation time series only. To realize dynamic characteristic extraction of the object time series [23], [24], we will estimate the behaviors of the hidden deterministic state variables X_i ($i = 1, 2$) of the actual observation time series. To estimate the deterministic rule, the reconstruction theorem by F. Takens [25], [26] is used.

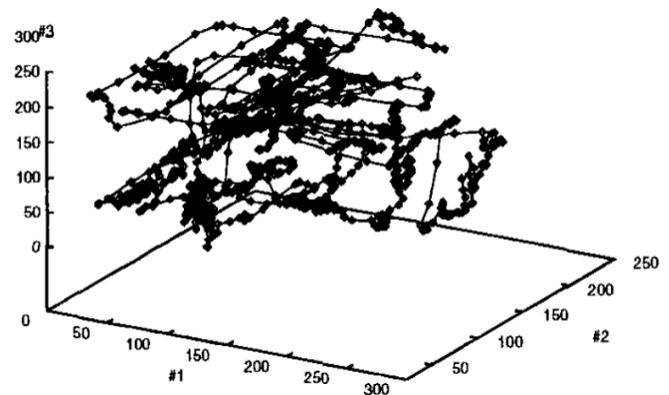


Fig. 1. Overview of the Trajectory of X_t .

A. Reconstruction Theorem of Takens

From two different scalar real observation time series ζ, η_i , $m_1 + m_2$ dimensional state variable,

$$X_1 = (\zeta_1, \dots, \zeta_1 - (m_1 - 1)\tau_1; \eta_1 - \beta, \dots, \eta_1 - (m_2 - 1)\tau_2 - \beta)$$

$$X_2 = (\zeta_2, \dots, \zeta_2 - (m_1 - 1)\tau_1; \eta_2 - \beta, \dots, \eta_2 - (m_2 - 1)\tau_2 - \beta)$$

$$X_l = (\zeta_l, \dots, \zeta_l - (m_1 - 1)\tau_1; \eta_l - \beta, \dots, \eta_l - (m_2 - 1)\tau_2 - \beta)$$

Are composed (T_1, τ_2, β : time delay). When analyzing based on actual observation data, how to choose the time delays τ_1, τ_2, β is a problem. Here, the parameters $m_1 + m_2$ are called embedded dimensions. $X_1, X_2, \dots, X, \dots$ are temporal orders, and the displacement from X_t, X_{t+1} at an arbitrary time t depends on the actual observation time series ζ_i, η_i .

By analyzing the trajectory of $m_1 + m_2$ dimensional state change X_t , We estimate the characteristics of the real observation time series ζ_i, η_i . As an example, the trajectory of the three-dimensional state variable X_t is shown in Fig. 1.

B. Judgment Method

In the space of the state variable X_t , El Niño Year / La Niña, the history of the year of Niña and non El Niño / non-La Niña. We think that it is different from the trajectory of the year.

III. EXPERIMENT

A. Data Used

The data used are SST (sea surface temperature) from the Japan Meteorological Agency and Make the monthly data of SOI (southern oscillation index). However, SST is the standard value of the monthly average sea surface temperature (30 years average value from 1961 to 1990).

SOI is the data of the difference between Tahiti and Oh on the South Pacific. It was indexed based on the atmospheric pressure deviation of Darwin on the Straia. It is a thing and is a measure of the strength of the trade wind, time limit. It is January 1972 to December 2001. Table I shows monthly mean sea level. Table I shows the mean values of SST per month derived from SST data from 1961 to 1990.

⁶ <https://www.worldscientific.com/doi/abs/10.1142/S0218127491000634>

⁷ https://en.wikipedia.org/wiki/Takens%27s_theorem

⁸ https://en.wikipedia.org/wiki/El_Niño-Southern_Oscillation

⁹ <http://www.sosmath.com/calculus/diff/der07/der07.html>

TABLE I. THE MEAN VALUES OF SST PER MONTH DERIVED FROM SST DATA FROM 1961 TO 1990

Month	Mean SST(deg.C)
January	25.4
February	26.2
March	26.9
April	27.1
May	26.6
June	26.1
July	25.2
August	24.6
September	24.6
October	24.6
November	24.6
December	24.9

It shows the reference value of water temperature. In addition, it was decided by the Meteorological Agency that the El Niño phenomenon occurred years is

- 1972-1973
- 1976-1977
- 1982-1983
- 1986-1988
- 1991-1993
- 1997-1998

It is judged that the La Niña phenomenon occurred years is

- 1972-1976
- 1984-1985
- 1988-1989
- 1998-2000

It is the 4th time of the year. These include El Niño phenomenon and there is a year when La Niña phenomenon occurred. In addition, one year is the period from January to December; also, in 1991-1993 El Niño phenomenon and La Nina in 1972-1976. There are two time periods for the El Niño and the La Niña phenomenon in 1998-2000.

B. Experimental Method

The proposed method analyzes the trajectory of mass points on multidimensional space, and from the viewpoint of visual understanding of the reader, this paper reports cases where $\beta = 0, m_1 = m_2 = 1$. That is, by setting $\beta = 0, m_1 = m_2 = 0$, and $\beta = 0, m_1 = m_2 = 1$, analysis on a two-dimensional plane can be performed. The reason for using the southern vibration index is that the sea surface temperature and the southern vibration index are interlocking with each other in order to take account of the influence.

A method for identification of El Niño / non El Niño is the followings:

1) The Meteorological Agency judged that El Niño phenomenon occurred. Appropriate reconstruction theorem for sea surface temperature and southern vibration index of year. Use it to display its trajectory.

2) The Japan Meteorological Agency did not judge that El Niño phenomenon occurred. Reconstruction for sea surface temperature and southern vibration index in a year. Apply the idea and display its trajectory.

3) The Meteorological Agency judged that the El Niño phenomenon occurred. The locus of the reconstruction space of the year and the Meteorological Agency issue the El Niño phenomenon. The difference from the trajectory of the reconstruction space of the year not determined to have been generated. Examination, El Niño passing the path of the year and non-El Niño. Find the area of the reconstruction space where the trajectory of the year does not pass.

4) In the locus of a certain year, the found reconstruction space. When that locus passes through the area of it is judged as El Niño year.

Identification of La Niña / Non La Niña can be done with the same method. The problems are the following:

1) Is there a region of reconstruction space through which the locus of the El Niño year passed and the trajectory of the non-El-Niño year does not pass?

2) Is there a region of reconstruction space through which the locus of La Niña passed and the trajectory of the non-La Niña year does not pass?

3) If the trajectory passes through the area of the found reconstruction space, does it agree with the judgment result by the definition of the Japan Meteorological Agency?

4) Does the decision timing by the proposed method become slower than the judgment timing based on the definition of the Japan Meteorological Agency?

C. Application of Takens Reconstruction

Below, the horizontal axis of Fig. 2 to 5 shows the difference from the standard value of monthly mean sea surface temperature and the vertical axis is the southern vibration index.

Fig. 2 shows the trajectory of (SST, SOI) from 1972 to 2001. The red line (72-73, 76-77, 82-83, 86-88, 91-93, 97-98) is the locus of the year determined by the Meteorological Agency that the El Niño phenomenon occurred and the green line (74-75, 78-81, 84-85, 89-90, 99-01). The blue line (94-96) is the locus of the year not judged by the Meteorological Agency that the El Niño phenomenon occurred (Fig. 3(a)).

Fig. 3(b) shows the locus of the year that was not judged by the Japan Meteorological Agency that the El Niño phenomenon occurred, and shows the locus of the year judged by the Japan Meteorological Agency that El Niño phenomenon occurred; it is expressed for each period.

Fig. 4 shows the trajectory of (SST, SOI) from 1972 to 2001. In the figure, the red line (72-76, 84-85, 88-89, 98-00) is the locus of the year determined by the Meteorological Agency that the La Niña phenomenon occurred and the green

line (86-87, 01), Blue line (77-83), Light blue line (90-97) is the locus of the year that was not judged by the Meteorological Agency that the La Niña phenomenon occurred. Fig. 5(a) shows the locus of the year determined by the Meteorological Agency that the La Niña phenomenon occurred for each period, and Fig. 5(b) shows that the La Niña phenomenon occurred and the Meteorological Agency In each period, the trajectory of the year that was not determined by the period.

As shown in Fig. 5(b), $SOI = 0$, so the state of $SST > 0.5$ becomes a short period and it is not judged as El Niño from the definition of the Japan Meteorological Agency, and from Fig. 5(b) $SOI < 0$, the state of $SST = -0.5$ became a short term and it is understood that there was a case that it was not judged as La Niña from the definition of the Japan Meteorological Agency. That is, the authors judge that the southern vibration index and sea surface temperature are necessary for early identification of El Niño / La Niña.

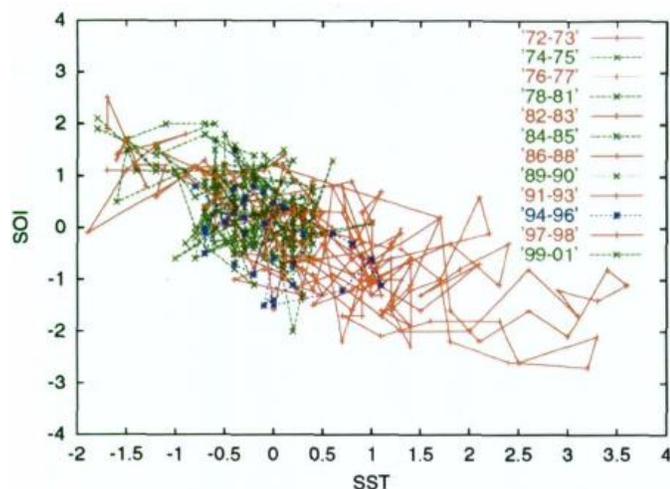
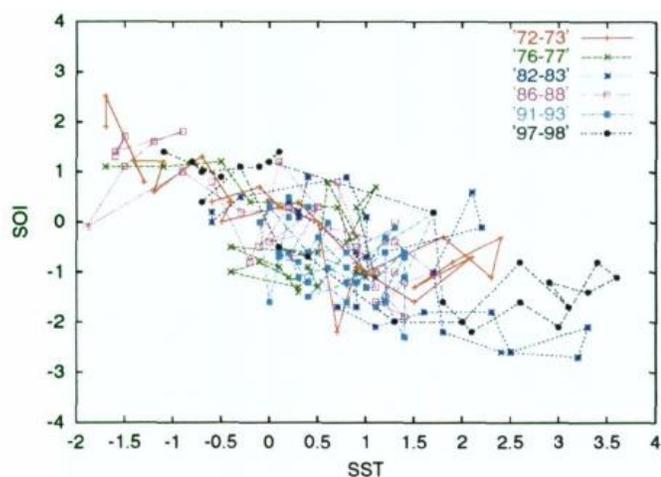
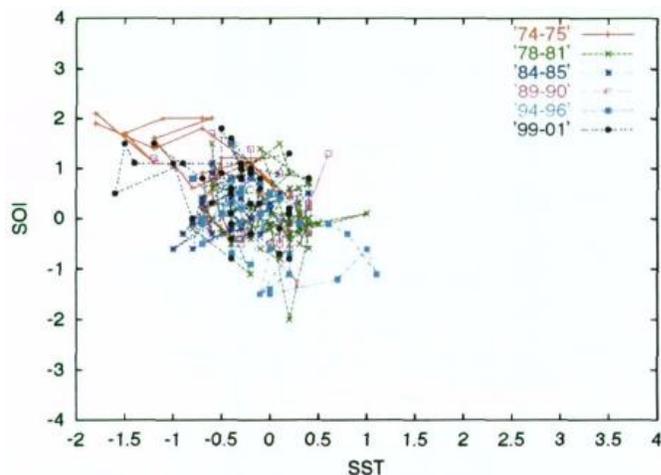


Fig. 2. The trajectory of SST-SOI Data during El Niño years and non El Niño years from 1972 to 2001.



(a) The trajectories of SST-SOI data during El Niño years.



(b) The Trajectories of SST-SOI Data during non El Niño Years.

Fig. 3. The trajectories of SST-SOI Data during El Niño years and non El Niño years.

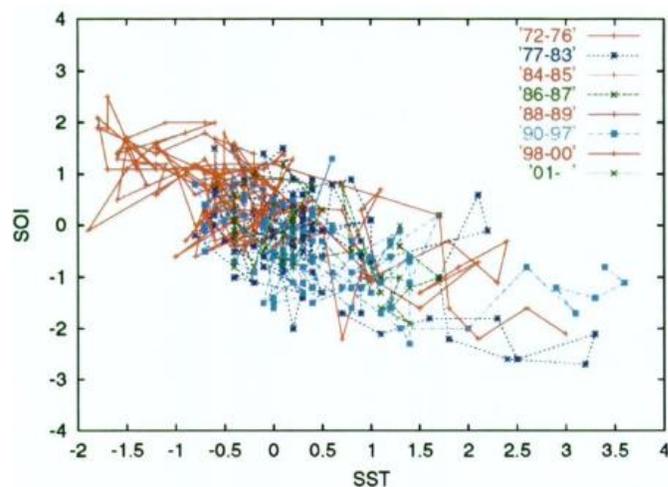
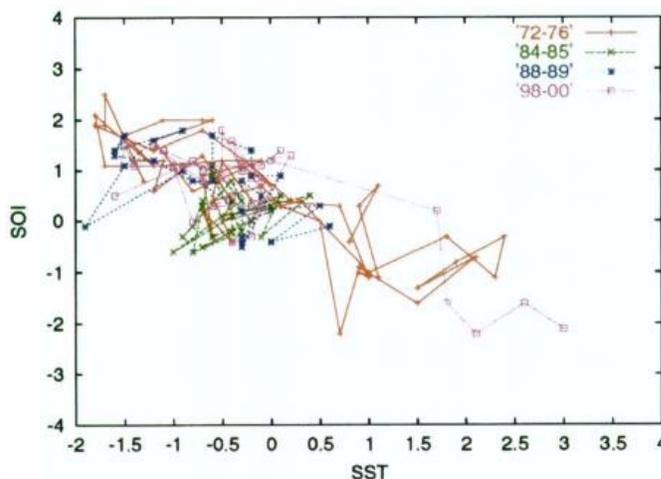
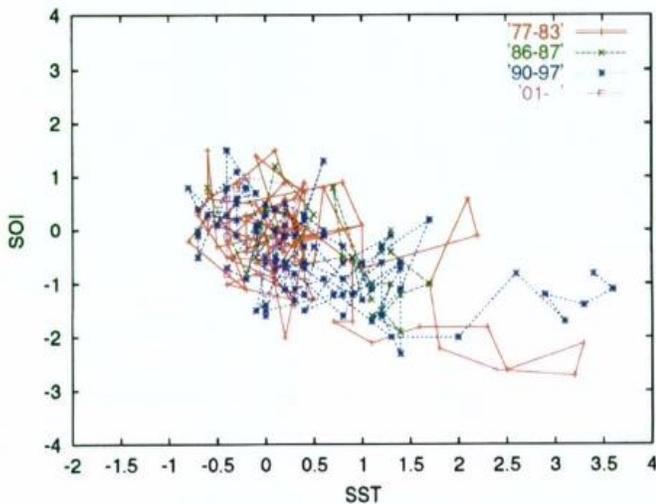


Fig. 4. The trajectory of SST-SOI Data during La Niña years and non La Niña years from 1972 to 2001.



(a) The trajectories of SST-SOI data during La Niña years.



(b) The trajectories of SST-SOI data during non La Nina years.

Fig. 5. The trajectories of SST-SOI data during La Nina years and non La Nina years.

Fig. 2 ~ Fig. 3 and Fig. 4 ~ Fig. 5, the trajectory of (SST, SOI) of the year judged by the Meteorological Agency that the El · Niño phenomenon / La · Niña phenomenon occurred and El · Niño phenomenon / It can be seen that there is a region different from the locus of (SST, SOI) of the year that was not judged by the Meteorological Agency when the La Niña phenomenon occurred. In other words, the locus of (SST, SOI) in the year that the Japan Meteorological Agency did not judge that the El · Niño phenomenon occurred does not pass through at least to the area of

$$SST > 1.1, SOI < 0 \quad (1)$$

It is understood that the locus of (SST, SOI) in the year that the Niña phenomenon occurred does not pass through at least to the region of

$$SST < -1.0, SOI > 0 \quad (2)$$

In addition, the locus of (SST, SOI) in the year when the Meteorological Agency judged that the El · Niño phenomenon / La Niña phenomenon occurred was passed, it was judged by the Meteorological Agency that the El Niño phenomenon / La Niña phenomenon occurred Once it enters an area where the trajectory of (SST, SOI) did not pass, it will be found that it will stay within that area for a while.

Tables II and III show the period of time required for judgment when identifying El Niño / La Niña using (1) and (2). From Tables II and III, it can be seen that the period required for judging the proposed method is shorter than the 10 months, which is the period required for judgment by El Niño / La Niña as defined by the Meteorological Agency. In other words, by using the definition of the Japan Meteorological Agency and the use of (1) and (2), it is possible to identify El Niño / La Niña at an earlier time. The advantage of using the definition of the Japan Meteorological Agency in combination is that in this paper $0.5 \leq SST \leq 1.1$. Table II shows the time period for identification of El Nino. Meanwhile, Table III, the time period for identification of La Nina.

TABLE II. THE TIME PERIOD FOR IDENTIFICATION OF EL NINO

Year	Proposed
1972-1973	4
1976-1977	Definition of the JMA
1982-1983	5
1986-1988	7
1991-1992	2
1992-1993	3
1997-1998	1

TABLE III. THE TIME PERIOD FOR IDENTIFICATION OF LA NINA

Year	Proposed
1972-1974	2
1974-1976	9
1984-1985	4
1988-1989	1
1998-1999	4
1999-2000	3

And analysis of the following region:

$$-1.0 < SST < -0.5 \quad (3)$$

can be considered. In other words, there is a possibility that the definition of the Meteorological Agency may be satisfied by continuing to stay in the area of

$$0.5 < SST < 1.1 \text{ or } -1.0 < SST < -0.5 \quad (4)$$

Indeed, in Table II, 1976 - 1977, it remained in the region of $0.5 \leq SST \leq 1.1$, which is an example satisfying the definition of the Japan Meteorological Agency.

IV. CONCLUSION

An identification method for earth observation data according to a chaotic behavior based on Takens reconstruction theory is proposed. The proposed method is examined by using the observed time series data of SST (Sea Surface Temperature) and the SOI (Southern Oscillation Index) data. The experimental results show that the time for the identification of the proposed method is not later than that of the existing method.

We confirmed that by using the definitions of the Japan Meteorological Agency and the use of (1) and (2). I can identify El Niño / La Niña at an earlier time. In other words, we do not necessarily need a numerical value for ten months by identifying the proposed method. We also confirmed that the time required for the identification judgment of the proposed method is about one month. The proposed method is not based on extrapolation method with numerical model or governing equation, but based on interpolation method using only actual observation time series.

This paper does not change the definition of the Meteorological Agency's El Niño phenomenon / La Niña phenomenon. By analyzing the boundary of the area where the

locus of El Niño Year (La Niña Year) passes and the trajectory of non-El-Niño year (non La Niña year) does not pass. El Niño / we confirmed that there is a discrimination method consistent with the identification result of La Niña.

Further investigation is required for validation of the proposed method with a variety of cases.

ACKNOWLEDGMENT

The authors would like to thank the fourth group of Information Science Department of Saga University for their useful discussions through the simulation studies.

REFERENCES

- [1] Nagasaka Koichi: "El Niño - Pacific Trade Wind and Extraordinary High Water Temperature in Eastern Equatorial Region," *Ships and Marine Weather*, 27, (1), pp. 4-8, 1983.
- [2] Yamagata Toshio: "A new wave of climate research - 8 - El Niño and Southern Oscillation," *Science*, 54, (11), pp. 699 - 705, 1984.
- [3] Tetsuzo Yasunari: "Global aspects of ENSO (El Niño / Southern Oscillation)," *Weather*, 33, (10), pp. 507-513, 1986.
- [4] T. Yasunari: "Global Structure of the El Niño / Southern Oscillation-Part I," *Meteorological Journal*, 65, (1), pp. 67-80, 1987.
- [5] T. Yasunari: "Global Structure of the El Niño / Southern Oscillation - Part II," *Meteorological Journal*, 65, (1), pp. 81-102, 1987.
- [6] T. Nitta and T. Motoki: "Abrupt Enhancement of Convective, Activity and Low-Level Westerly Burst during the Onset Phase of the 1986-87 El Niño," *Meteorological Journal*, 65, (3), pp. 497-506, 1987.
- [7] T. Maruyama and Y. Tsuneoka: "Anomalous Short Duration of the Easterly Wind Phase of the QBO at 50 hPa in 1987 and Its Relationship," *Meteorological Journal*, 66, (4), pp. 629 -634, 1988.
- [8] T. Nitta: "Development of a Twin Cyclone and Westerly, Bursts during the Initial Phase of the 1986-87 El Niño," *Meteorological Journal*, 67, (4), pp. 677-681, 1989.
- [9] Hagiwara Samurai: "Efforts to solve the El Niño phenomenon," *Prometheus*, 11, (4), pp. 58 - 60, 1987.
- [10] Masahide Kimoto: "El Niño and prediction of extreme weather," *Science*, 65, (6), pp. 389-397, 1995.
- [11] Koshiba Atsuto: "Statistical features of temperature and precipitation in the world during the El Niño La Niña phenomenon period," *Journal of the Japan Atomic Energy Research Timetable*, 49, (5), pp. 143-149, 1998.
- [12] Kazunobu Nakamura, Yoshinobu Nikaido, Ikuo Yoshikawa, Naoyuki Hasegawa, Tadashi Ishii, Good: "Toward the prediction of the El Niño phenomenon," *Weather Timetable*, 65, pp. S 39 - 85, 1998,
- [13] Koji Wada, Hiroyuki Kato: "The current state of research on prediction of El Niño and related abnormal weather," *Central Research Institute of Electric Power Industry Research Report T, 98056*, pp. 1 - 23, 1999.
- [14] Meteorological Agency Climate Oceanic Weather Department Climate Information Division: "Monitoring and prediction of global warming and El Niño phenomenon," *river*, 57, (12), pp. 19-24, 2001.
- [15] Kazuo Kurihara: "Use of TAO / TRITON buoy data for monitoring and prediction of El Niño phenomenon," *Weather Timetable*, 69, pp. S47 - S54, 2002.
- [16] Kubota Masahisa: "A study on the interaction between the atmosphere and the ocean using satellite observations and numerical models," *Weather*, 49, (5), pp. 369 - 384, 2002.
- [17] Tatsunori Matsuura, Michiaki Yumoto, Satoshi Izuka, Hisashi Yoneya: "Simulation of Typhoon and El Niño Southern Oscillation by High Resolution Atmosphere and Ocean Binding Model," *Journal of the Institute of Electrical Engineers of Japan*, 912, pp. 36-42, 1999.
- [18] Takabuki (Nakagome) margin: "Large scale meteorological phenomena in tropical zone and TRMM observation - 1997-98 El Niño's termination and Madden-Julian vibration," *Ocean*, 31, (6), pp. 383-390, 1999.
- [19] Saeki Riro · Takahashi Yonobu: "About the work of the El Niño Monitoring Center," *Weather Timetable*, 59, (4), pp.161-170, 1992.
- [20] Tadashi Ando: "El Niño Surveillance Center," *Ocean Research*, 2, (2), pp. 109-115, 1993.
- [21] Meteorological Agency El Niño Observation Forecast Center: "El Niño events occurred," *Yuki*, 29, pp. 80 - 85, 1997.
- [22] Fujiwara Sachiko: "About the updated" El Niño surveillance bulletin ", *weather*, 519, pp. 16736-16739, 2000.
- [23] Masahide Kimoto: "El Niño Chaos," *Mathematical Sciences*, 34, (11), pp. 81-85, 1996.
- [24] J. D. Neelin and M. Latif: "Numerical modeling for prediction El Niño dynamics," *Parity*, 14, (10), pp. 15 - 21, 1999.
- [25] F. Takens: "Detecting Strange Attractors in Turbulence," in *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, Springer, 1981.
- [26] L. Noakes: "The Takens embedding theorem," *International Journal of Bifurcation and Chaos*, 1, (4), pp. 867-872, 1991.
- [27] Kohei Arai, William Emery, El Niño phenomena analysis with Earth Observation Satellite data, *Global Observation International Network, Demonstration and Symposium*, Tokyo, June 1995.
- [28] Kohei Arai, Kaname Seto, El Niño and La Niña Discrimination Based on Takens Reconstruction Theory, *Journal of Remote Sensing Society of Japan*, Vol.23, No.2, pp.157-163, 2003.
- [29] Kohei Arai, Prediction method of El Niño Southern Oscillation event by means of wavelet based data compression with appropriate support length of base function, *International Journal of Advanced Research in Artificial Intelligence*, 2, 8, 16-20, 2013.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. <http://teagis.ip.is.saga-u.ac.jp/index.html>

Adaptive Return of e-Training (ROT) based on Communication Technology

Fahad Alotaibi

Department of Information System
Faculty of Computing and Information Technology,
King Abdulaziz University,
Jeddah, Saudi Arabia

Abstract—Persistent economic insecurity and harsh severity actions across the world push businesses either to cut down on training costs or to be very painstaking in choosing a training program that conveys palpable outcomes in a short period of time. Nevertheless, in most cases businesses are still unable to reckon Return of e-Training (ROT) in advance for better allocation of training budget and decision on a proper training plan in line with the business policy. The purpose of this paper is to appraise the practical worth of the applicability and usability of the Adaptive ROT in the enterprises with a particular regard to evaluating the impact of e-training in companies. A case study of gauging the profit of e-training in the Blackboard systems has been conducted. The outcome of this study is judged to be positive, given the efficacy of the Adaptive ROT Evaluation Model for e-training in companies.

Keywords—Return of e-Training (ROT); evaluation models; blackboard; e-learning; Key Performance Indicator (KPI)

I. INTRODUCTION

Smart Learning and Development (L&D) groups are busy defining the factors that people consider in their prioritizations, by demonstrating the prolific effort of their labor force [1]. Motivating workforces to make use of technology is a big controversial issue, given their weak enthusiasm to get familiar with L&D for learning [6], [12]. Actually, they want to better their job-related skill sets to improve their career prospects. In this framework, this inclination to better their individual performance through the use of technology is called “learning” [8], [22]. Identifying this drive in the employee is vital to teaching technology, by guaranteeing micro (individual goal) and macro (organizational goal) development. Although companies are aware of the significance of training of their personnel, they still calculate the financial gains too.

Organizations downturn forced management to analyze the profit of training, taking into account the financial portions within their monetary restraint [5], [7], [8], [10]. Skill sets have sufficiently proven that the adoption of technology and training can function properly if businesses ascertain that they can make profits in case they opt for training and accept as true that novelty in tools of assessment and processes is unavoidable. A wide-ranging and inexpensive ROT system must adopt business strategic practices. By developing an ROT system that is reasonably priced and comprehensive, and

is an outcome of strategic business planning, the quality and efficacy of training can increase [10], [23]-[26].

This emphasis on ROT denotes the new-found focus put on the improvement of the professional practices by internal and external trainers.

II. METHODOLOGY

In this study, the methodology of adaptive ROT system, based on collaboration and use of communication technology, has a block diagram as shown in Fig. 1.

The study consists of four (4) major sections: 1) Measuring the training by using KPI, 2) Adapting employees with e-learning, 3) Measuring ROT, and 4) Measuring the time period of ROT.

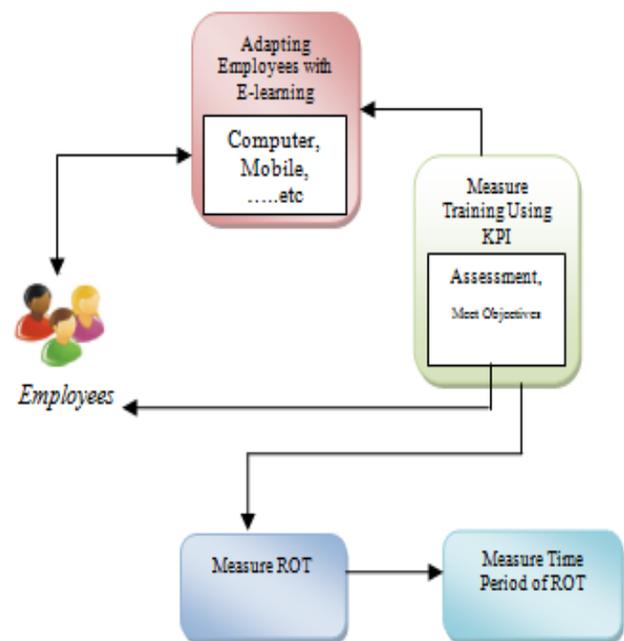


Fig. 1. The framework of adaptive ROT.

This Research Project is sponsored by the Deanship of Scientific Research (DSR)-King Abdulaziz University –Jeddah-Saudi Arabia

III. MEASURING AND BENCHMARKING TRAINING PROGRAMS OUTCOMES USING KPI

Management usually has unclear understanding of ROT, and, quite often, it is not conscious of training costs. For this reason, it is not surprising that educational programs become the first victim of cost-reduction campaigns when a business faces a depression.

It follows then that training and organization development manager has to take the following actions to make sure that ROT will measure the employee's performance [8], based on managing by objective system for the company: (1) He has to base the policy of training needs assessment on the following necessary factors: (1.1) The annual performance results based on 90 degrees system at least, (1.2) an updated Job description, (1.3) The updated changes in the annual overall operation plan for the company, and (1.4) the results of suggestions system. (2) He has to design the training courses and outlines in order to cover the reality means, depending on the cost center policy for each department, which forces the department manager to ponder on what exact needs for the department to achieve within the estimated yearly expenses. (3) He has to divide the training activities between the external training and internal training (on-job training) to guarantee that the following things take place: (3.1) First, the business needs to gain new experience from the external training to its personnel, (3.2) Then, they need to use internal training and On-the-Job Training (OJT) to make them transfer what they learned from the external training to their colleagues (to make sure that all are on the same page and none has a rare experience). (4) Before every training course, the business has to write a report on the training needs and hand it to the instructor; regularly assess (together with the instructor) the way of implementing the training in the real life by a final daily training report; and the trainees have to sign with knowledge. (5) One month after the end of the each course, the training manager should visit the department which got the course, consider whether they implemented the training by the signed report or not, and measure the ROT.

IV. ADAPTING EMPLOYEES FOR E-LEARNING TECHNOLOGY

Familiarizing the employees with e-learning technologies can be achieved by following five tips as stated in [9], [13], [26]. These tips can work for peers in other organizations and help to accomplish the L&D objectives, by utilizing the appropriate technologies. Below are these tips:

- **Address Your Employee's Biggest Work Challenges:** One should collect his employees' specific work challenges. What to do with them will be explained in the 4th tip, but to summarize, one can say that the adoption of the right technology and approach means one can specifically address these challenges without any problem of technology use at all. Client experiences are becoming more personalized—deviating from standardization—as they are facilitated by smart technology, and consumer expectations are also becoming associated with corporate expectations.
- **Don't Dress up HR and L&D Priorities as Employee Priorities:** On a common-sense level, managers should

have access to it. But what are the company primacies that are inhibiting the managers from having access to it? Their world is intricate and challenging, and so considering what could be seen as supplementary (non-business critical) activities as priority is unrealistic. Very often, an incongruity of primacies (between HR and employees themselves) originates from interest conflicts. So, one should get close to them and comfort them with their insistent distresses (their 'what?')—and he might just find ways to impact the 'how?'

- **Share Career Stories of How others Have Progressed:** One should gather as many job stories as he can of individuals in his company across diverse ranks, functional disciplines, and development in the organization. Host panel events to enhance his video stories, and thus participate in dialogues and bring them to life.
- **Use 'Resources' to Keep Employees in the Workflow:** The objective is to provide just enough perception, instruction, or information to aid the workers to advance with their work, with the addition of more self-reliance and proficiency than they would without it. This would be done with the appropriate technology, such as Loop (which is goal-oriented for your assets). Bear in mind that the user's know-how is as important as the presented content. If a resource can't be accessed on-demand, on-the-go, and as simply as a web-search, then Google would win! The appropriate tools make all the variance. Moreover, numerical resources can be created and shared in seconds, with Loop. So, one should not cut corners; he should rather spend on the suitable utensils.
- **Run Campaigns to Drive Traffic and Repeat Visits:** Unluckily, we don't live in 'Field of Dreams' (or Wayne's World 2) and if we construct that world, dreams will not come. We have to conduct operations to trade the worth of your assets and guarantee unrelenting commitment. Even YouTube has weekly summaries of their most watched videos in order to sustain commitment. One's own weekly summaries can be produced and shared for the sake of energy circulation and repeated visits; and one can also generate worker stories that prove the worth of committing time to his resources through the upshots that can be attained. One has to be insightful and make use of the best weekly summaries out there, just because his content will not do this on its own. He can achieve that in an easy way by showing what's popular and showing what's new.

A. Advantages of e-Learning

Many companies have made investments in e-Learning over the last twenty years, principally for the following advantages:

- Self-paced, interactive, and more appealing learning (from learner perspective).
- Access at any time, and from anywhere (on-demand availability).

- Cost-effective (particularly when the training is delivered to a large audience).
- Less troublesome conveyance (in contrast to ILT).
- Easy tracing of learner's advancement and accomplishment (from business perspective).
- Message-consistency and easy content- updating.

B. Measuring Return on Investment (ROI) of Online Training

The adoption of e-learning is gaining further impetus, because traditional e-Learning moves towards mobile learning or m-Learning and provides learners with the flexibility of learning on the device of their choice (notably tablets and smart phones) [18]-[22]. E-learning and m-learning provide several paybacks to establishments. Nonetheless, the emphasis is now shifting to determining its influence and the Return On Investment (ROI) of online training. A successful e-learning inventiveness must be capable of bringing gains that are more than the expenses [17]. ROI is the return on investment that a business achieves, and can be calculated as:

$$ROI = \text{Gain or Return} / \text{Cost} \quad (1)$$

It may be calculated by way of two parameters, specifically the expenses paid out (or charge suffered) and the Worth/Achievement amassed (or return).

Calculating Return on Training Investment (ROTI): Return on Investment (ROI) is the correlation between monetary paybacks acquired from something (in this case a learning program) and the overall budget of that thing. The objective of an ROI scrutiny is mostly to assess whether the profits are greater than the expenses, i.e. to perceive whether the expenses were worth it. It is worth noting that ROI of learning can only be measured reliably by means of Training Check, if the conditions below are fulfilled: (1) Availability of trustworthy information on variations to related organization performance measures (to be reasonably estimated by key stakeholders). (2) Possibility of assigning financial value Changes to the selected performance measures [8], [22]. (3) Identification of the expenses related to the development, delivery and management of the learning. If these conditions are fulfilled, one can use Training Check to prepare a Return-on-Training Investment (ROTI) report, by means of the ROTI Calculator function, and by following the stages below. Once the relevant financial data have been entered, the Calculator will routinely compute the ROTI % such as below:

$$ROTI = \frac{\text{£ BENEFITS} - \text{£ COSTS}}{\text{£ COSTS}} \times 100 \% \quad (2)$$

The Benefit to Cost ratio will also be calculated as follows:

£ BENEFITS : £ COSTS, as stated in [2-5]. There is also another possibility to determine the 'Payback Period', i.e. the time it takes to reimburse the expenses. Remarks on ROTI Outcomes: It is worthwhile to get primary arrangement from learning program supporters regarding the objective level of ROTI. As a customary norm, ROTI levels beneath about 20% are generally assessed to be low. In reality, however, it is reasonably customary for ROTI ratio records to be very high, example 500% or more. Although elevated ROTI records may

amaze high-ranking executives, they possibly will also nurture suspicions, specifically amongst those who are commonly unimpressed by the significance of on-site-training.

For this reason, it is significant to integrate ROTI computations with other aspects of assessment. Furthermore, demonstrating a constant correlation between the training and advances in career and business accomplishment will boost substantial reliability to ROTI records. Likewise, wherever the ROTI records attained are truncated or undesirable, reaction from other assessment ranks can be used to assist identify any hindering causes.

C. Calculating Return on Training Investment (ROTI)

Computing the return on investment from a training program could be significant once the training program is seen as a substantial deal by the organization, or once it is brought into line with the accomplishment of a particular planned or real goal. It could as well be worthwhile once it is unclear if a program will engender any economic returns or what those returns might rise to.

Nonetheless, despite the fact that ROTI might have a significant role in a training program assessment, an ROTI measurement only will not customarily be enough to commend the company situation for a training program or convince high-ranking executives to act in a specific manner. Because of this that very frequently it is only one minor component of the worth of the training. Contingent with the approved goals and anticipations of the training program, factors such as ROTI are very often calculated as follows:

On a regular basis, computing the ROTI from a learning program should only be instigated when the following criteria are met (i.e. in case not all these criteria are met, it follows then that one has to earnestly reflect on if it would be advisable to invest time, energy and assets on making an ROTI investigation):

- The existence of substantial monetary overheads that the learning program requires.
- The ROTI analysis must be meaningful / important to the program's sponsors.
- The training objectives must be plainly well-defined and their attainment must be susceptible to influence on places of the premeditated or effective significance.
- The availability of information on pertinent changes to performance.
- The existence of sufficient trainees to influence the company achievement and draw economic advantages.
- The trainees should be allowed worthy chances to implement their training to the place of work.
- Identification of direct and indirect costs of training.
- Attribution of credible financial values by the main investors to changes to performance (see Note 1 below).

- Isolation of the training factors from other causes and allocation of the monetary aids in view of that (see Note 2 below).

For instance, if advantages in personnel preservation have yielded an economic profit of £5,000 to the business, and it is predictable that the learning is accountable for 50% of the variation in preservation (and the other 50% being accredited to other causes), it follows then that the total economic profit associated with the learning is computed as £2,500 (i.e. £5000 x 50%). This total is then used as component of the computation of the Return on Training Investment.

Participants need to be motivated to stay on the conventional side at the time they make budgetary approximations. If these estimations are irrationally high, this may harm the reliability of the ultimate ROTI facts.

D. Measuring Time Period ROTI

There exist no stable timespan over which one may determine the ROI of a training program [23]-[25]. Some generally utilized instances encompass three points: (1) From three months to twelve months after training has been completed (allows time for the transmit of training to the jobsite). (2) One financial year (audit period)/the period of a product cycle. (3) Average period of target audience employees retention in the company.

TABLE I. AN EXAMPLE OF PAYBACK PERIOD CALCULATION

Number of months over which benefits are calculated	12
Total Benefits	81,500
Monthly benefits =	6,792
Total Costs	15,000
Payback period	2.2 months

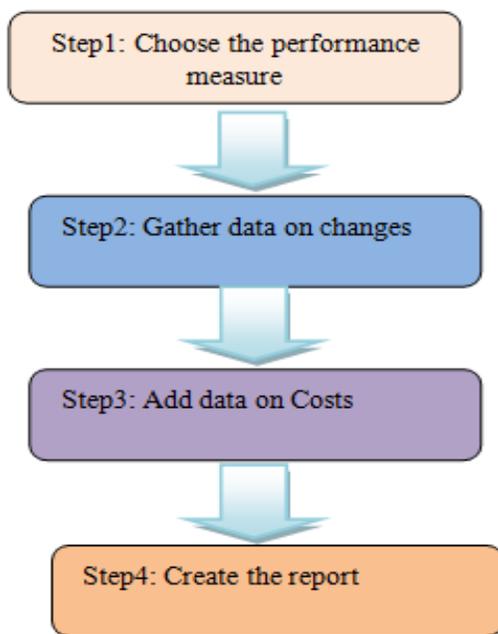


Fig. 2. Stages for collecting ROTI data and creating report.

Definition 1: Payback Period: The Payback Period is the time spent to recompense back the expenses, i.e. when the incurred expenses equal the accrued profits [2]-[6], [14]. A small return timespan is expected to amaze and may add to the company situation for spending on additional learning. The ROTI Calculator computes the return timespan by matching stated paybacks with stated expenses such as indicated in Table I:

$$\text{Payback Period} = \frac{\text{COSTS}}{\text{monthly benefits}} \quad (3)$$

Gathering ROTI data and creating the ROTI report commonly goes through four fundamental stages, these stages can as presented in Fig. 2 and can be delineated as follows:

Stage 1: The choice of the performance measures to be used. First, one has to work with the main sponsors to select the appropriate corporate performance measures which are meant to be used as a foundation for the ROTI scrutiny. Samples of measures comprise variations in: (1) Client contentment and retention rates / degree of customer dissatisfaction. (2) Output/productivity amounts / sales volumes / worker income amounts. (3) Total of monthly sick-absence days / total of annulled training days/sessions. (4) Depletion rates / non-compliance/ annual misfortunes rate / staffing expenses.

It is essential that the selected measures have the following features: They should be quantifiable, or can be rationally estimated by crucial participants, they should be assigned a financial value by investors, and finally, they should be pertinent, i.e. variations in them should be associated with the learning. (Note: When parameters other than the learning that may impact variations to the chosen measures exist, participants have to approximate the ratio of variation which may be openly associated with each parameter).

Stage 2: Data Collection on variations: Next, one has to collect information on the chosen accomplishment measures. He needs to gather the ensuing data like: Monetary information associated with the variations in accomplishment, and Approximations of the % impact of the learning on these variations comparative to other potential powers. Let us assume the following parameters:

$$\text{Total Financial Benefit of Performance Change} = TFBPC,$$

$$\text{Influence of the Training} = IOT, \text{ and}$$

Total Financial Benefit Attributable to the Training = TFBAT then TFBAT can be calculated as:

$$TFBAT = TFBPC \times \%IOT \quad (4)$$

For instance, if the overall economic profit of perfections in employees' maintenance is £10,000, and the learning is judged to have been accountable for 50% of that variation, then we obtain what follows:

$$TFBAT = £10,000 \times 50\% = £5,000 \quad (5)$$

This total (i.e. £5,000) will be employed as constituent of the ultimate ROTI computation.

Stage 3: The addition of information on expenses: Once the economic paybacks from the learning program have been added, then data of the expenses connected with the learning will need to be entered. To start this procedure, we need to click the 'Add Data on Costs' key on the ROTI Calculator page. This will trigger the 'Cost Calculator'. After that we only enter the data of costs as prompted. Expenses connected with a learning program would fit in the comprehensive types underneath: (1) Running / development expenses (e.g. developer costs, design, printing). (2) Conveyance expenses (example, facilitator payments, venue, training resources) / turnout expenses (example, worker-discharge overheads, travel, lodging).

Stage 4: Creation of the report: Once we have entered all of the data on monetary profits and expenses connected with a learning program into the ROTI Calculator, we will be able to write the ROTI report. We simply have to press the 'Create Report' key on the final page of the Cost Calculator. The report has to contain a précis of the paybacks and expenses added, and offer a scrutiny of the ROTI %, the Benefit to Cost ratio, and the Payback Period. ROTI Reports are put in storage on the My Reports page and can be edited there.

V. RESULTS AND DISCUSSION

The assessment of performance in our analysis is measured in terms of the learning results to the enterprise and performance objectives, the adoption of the right learning approach to administer the virtual trainings, the implementation of an effective evaluation approach [7], [11], [17], the cover pushing to knowledge execution, the provision of a podium for cooperation in training (social learning), and the obtainment of user feedback and its use to update our approach. A case study of gauging the profit of electronic training in the Blackboard systems has been conducted in this analysis. Table II displays the assessment of the return of electronic training in the Blackboard systems.

Gauging the return of electronic training in the Blackboard systems can be sketched as follows:

It offers great chances for learners to interact with the syllabus outside the lecture room anywhere and at anytime through this electronic system, which provides the learners with an array of tools to view the content of the scientific resources and interact with them in many methods, in addition to communicating with the instructor and the rest of the learners enrolled in the same course by different electronic devices. It comprises tools and means that empower institution participants to develop dynamic and interactive courses with ease, while managing the content of these courses in a simple and simple manner, so that they can accomplish the daily tasks of the learning process successfully.

This system permits the trainer to develop cohesive electronic courses, make notes / outlines of the material / the required jobs and advertisements, and allows the lecturer to present periodical works, and personally original examinations and results.

TABLE II. THE EVALUATION OF THE RETURN OF ELECTRONIC TRAINING IN THE BLACKBOARD SYSTEMS

Investment	Calculations
Cost of course	€ 20,000.00
Number of students	20
Cost per student	€ 1,000.00
Return - Time Savings	
Average total cost of employee per year	€ 45,000.00
Average total cost per hour	€ 25.00
Period of improved performance	12
Value of time saved	€ 894.60
Return - Staff Turnover Savings	
Average cost of recruitment & induction	€ 6,750.00
Average cost of other training & warm-up period	€ 5,625.00
Impact on staff turnover as proportion of all benefits	10%
Value of reduced staff turnover	€ 1,237.50
Summary	
Total Investment per student	€ 1,000.00
Total Return	€ 2,132.10
ROI %	113%

The system allows the trainer to make a review of lessons, to remotely conduct assignments and electronic tests for training, or to use the institution computer lab, as well to manage discussions and inquiries with learners, or between them. The instructor can design electronic modules with the incorporation of multimedia and diffusion to learners remotely.

The system offers the advantage of downloading the material for the learner, so that he can follow the learning offline, and enables the instructor to put the curriculum on CDs, and enables the learner to review the material through the mobile phone (Pocket PC). The use of the Blackboard 9.0 learning management system makes it easier for an instructor to craft an online course site without the experience of software development and all the necessary access to the system, and the use of the mouse (point-click) to build its decision and follow the sequential guidelines offered by the system. Therefore, with a little training, organization participants can easily build interactive courses and add many features to enhance the course supplied by the system.

The system offers many tools to accomplish the learning process through the integration of the Blackboard and WebCT systems and the release of Blackboard 9.0, which is powerful and highly efficient in terms of ease of use and training management. This program is also used in the universities (King Abdul-Aziz - Electronic - Princess Nora- AlDammam).

REFERENCES

The Blackboard is an information system for the education management system, and it offers follow-up of learners and observing the efficiency of the educational process in the educational organizations. It entails tools and means that arm institution participants with the ability to build dynamic and interactive decisions very easily, with the management of the content of these courses in a supple and simple manner, and to carry out the daily tasks of the educational process successfully. It can present periodical work, examinations and results on a well-timed basis, and evaluations, whether in the interim or final tests [12], [13], [15], [16].

This system allows direct communication with learners through discussion windows and focused and generalized e-mails. This system can be linked to other electronic learning systems, and interact with these systems in an integrated manner. It allows the possibility of using the Internet Mail with the possibility of placing files attached to the mail. The system involves the existence of the bulletin board that supports the mathematical symbols, images and PowerPoint files, with the ability of the system to archive these things

VI. CONCLUSION AND RECOMMENDATIONS

The implementation of an effective evaluation approach helps us to assess whether the learning satisfied the necessary mental degree and was truly able to weld the recognized breach. In this research work, we have proposed an Adaptive ROT, a case study of e-training in the Blackboard systems (for King Abdul-Aziz - Electronic - Princess Nora- AlDammam) has been conducted. We have outlined a worthy method to achieve the efficient return of e-training in the Blackboard that can supplement or complement the e-learning package. Other recommendations that can be considered in achieving efficient return of e-training are by providing a podium for cooperation in training (social learning), and investigation confirms that approximately 20% of our knowledge ensues from feedback and from watching our workmates in action (mates, seniors, or role models). It is worth noting that just 10% of knowledge results from official learning. Offering forums for social or casual education will be conducive to learning and can also be utilized to generate real-life stories of accomplishment. We can track down user feedback and use it to bring up-to-date our strategy. In the course of the online progress, we have to gather feedback from target learner groups. This needs to be done as we progress.

ACKNOWLEDGMENT

This paper was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University. The authors, therefore, acknowledge with thanks to DSR's technical and financial support.

- [1] Barrett, A & O'Connell, P.J. (2003) Does Training Generally Work? The Returns on In-Company Training, ESRI: Dublin.
- [2] Anderson, SB and Ball, S. The Profession and Practice of Program Evaluation, Jossey Bass, San Francisco 1978.
- [3] Anderson, V. Value and Evaluation: from return on investment to return on expectation. Report by University of Portsmouth Business School to CIPD, November 2007.
- [4] Bersin, J. The Training Measurement Book: Best Practices, Proven Methodologies, and Practical Approaches. Pfeiffer, 2008.
- [5] Birnbrauer, H. (1987). Evaluation techniques that work. Training and Development Journal, July, 53-55.
- [6] Bowsher, J. (1990). Making the call on the COE. Training and Development Journal, May, 65-66.
- [7] Brinkerhoff, R. O. (1988). An integrated evaluation model for HRD. Training and Development Journal, February, 66-68.
- [8] Bumpass, S. & Wade, D. (1990). Measuring participant performance - An alternative. Australian Journal of Educational Technology, 6(2), 99-107.
- [9] Bushnell, D. S. (1990). Input, process, output: A model for evaluating training. Training and Development Journal, March, 41-43.
- [10] Erkut, S. & Fields, J. P. (1987). Focus groups to the rescue. Training and Development Journal, October, 74-76.
- [11] Foxon, M. (1989). Evaluation of training and development programs: A review of the literature. Australian Journal of Educational Technology. 5(1), 89-104.
- [12] Hewitt, B. (1989). Evaluation a personal perspective. Training and Development in Australia, 16(3), 23-24.
- [13] Kirkpatrick, Donald L. (1994). Evaluating Training Programs: The Four Levels. San Francisco: Berrett-Koehler Publishers.
- [14] Lombardo, C. A. (1989). Do the benefits of training justify the costs? Training and Development Journal, December, 60-64.
- [15] Newstrom, J. W. (1987). Confronting anomalies in evaluation. Training and Development Journal, July, 56-60.
- [16] O'Donnell, J. M. (1988). Focus groups: A habit-forming evaluation technique. Training and Development Journal, July, 71-73.
- [17] Poulet, R. & Moullet, G. (1987). Putting values into evaluation. Training and Development Journal, July, 62-66.
- [18] Phillips, Jack J. (1994). Measuring Return on Investment: Volume I. Alexandria, VA: American Society for Training and Development.
- [19] Phillips, Jack J. (1996). "ROI: The Search for Best Practices." Training & Development 50 (February) 2:42-47.
- [20] Phillips, Jack J. (1997a). Handbook of Training Evaluation. Third Edition. Houston, TX: Gulf Publishing.
- [21] Phillips, Jack J. (1997b). Measuring Return on Investment: Volume 2. Alexandria, VA: American Society for Training and Development.
- [22] Phillips, Jack J. (1997c). Return on Investment in Training and Performance Improvement Programs. Houston, TX: Gulf Publishing.
- [23] Robinson, Dana Gaines and J.C. Robinson. (1989). Training for Impact: How to Link Training to Business Needs and Measure the Results. San Francisco : Jossey-Bass Publishers.
- [24] Senge, Peter M. (1990). The Fifth Discipline: The Art and Practice of the Learning Organization. New York, NY: Currency Doubleday. Training. (1996). Industry Report. Vol. 33, no. 10: 36-79.
- [25] Weatherby, N. L. & Gorosh, M. E. (1989). Rapid response with spreadsheets. Training and Development Journal, September, 75-79.
- [26] Wigley, J. (1988). Evaluating training: Critical issues. Training and Development, 15(3), 21-24.

Comparison of Hash Function Algorithms Against Attacks: A Review

Ali Maetouq, Salwani Mohd Daud, Noor Azurati Ahmad, Nurazeen Maarop, Nilam Nur Amir Sjarif, Hafiza Abas

Advanced Informatics Department
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia

Abstract—Hash functions are considered key components of nearly all cryptographic protocols, as well as of many security applications such as message authentication codes, data integrity, password storage, and random number generation. Many hash function algorithms have been proposed in order to ensure authentication and integrity of the data, including MD5, SHA-1, SHA-2, SHA-3 and RIPEMD. This paper involves an overview of these standard algorithms, and also provides a focus on their limitations against common attacks. These study shows that these standard hash function algorithms suffer collision attacks and time inefficiency. Other types of hash functions are also highlighted in comparison with the standard hash function algorithm in performing the resistance against common attacks. It shows that these algorithms are still weak to resist against collision attacks.

Keywords—Hash function algorithms; MD5; PRIMEDS160; SHA-1; SHA-2; SHA-3

I. INTRODUCTION

Among the most useful primitives that are crucial for data security is the cryptographic hash function, which offers message authentication, data integrity, and digital signature [1]-[3]. Additionally, it is employed as a core element of cryptographic protocols, secure transactions and cryptocurrencies. Fig. 1 presents an output of a fixed length (termed as a message digest or hash code) that uses a one-way function (known as a hash function) with an input of arbitrary length (also termed as a “message” or “plain text”) [4].

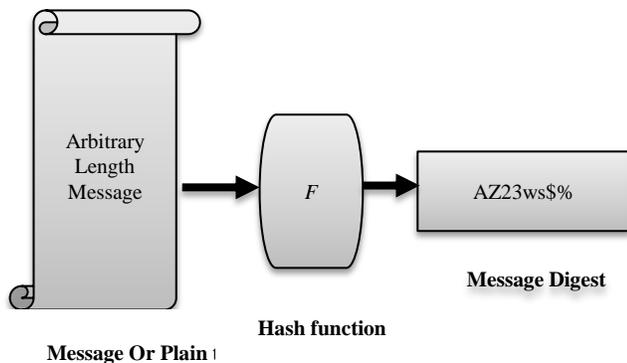


Fig. 1. Hash function.

The mathematical definition of a hash function (H) is defined as follows:

$$H: \{0, 1\}^* \rightarrow \{0, 1\}^n \quad (1)$$

Where, $\{0, 1\}^*$ refers to the set of binary elements of any length including the empty string. Meanwhile, $\{0, 1\}^n$ is used to refer to a set of binary elements with length n . Thus, a set of fixed-length binary elements is mapped to arbitrary-length binary elements using the hash function.

The organization of the paper is as follows. In Sections II and III, the basic concepts such as security properties and applications of hash functions are discussed. A literature review on the most popular hash function algorithms is provided in Section IV. Then, the comparison of the standard hash algorithm based on the general properties and common attacks are discussed in Section V. Many researchers have also proposed their own algorithms as discussed in Section VI.

II. PROPERTIES OF HASH FUNCTIONS

Several properties of security must be satisfied for cryptographic hash functions [5], [6].

A. Resistance to Collision Attacks

It would be impossible for the attacker to find the same hash value or $H(M)$ for two messages (M, M'). A collision attack happens when a pair of distinct messages having the same hash as shown in Fig. 2. The hash function must have the property of not producing same hash value for different messages.

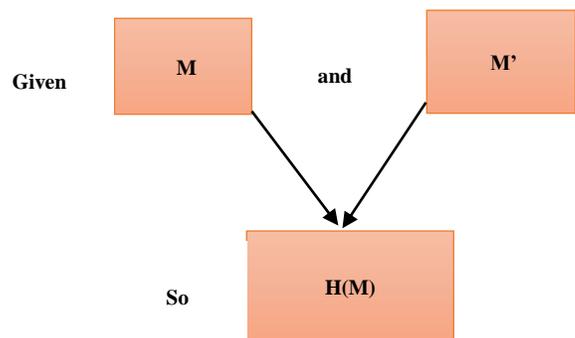


Fig. 2. Collision resistance.

B. Resistance to Pre-Image Attacks

A preimage is a message that hashes to a given value. In a preimage attack, it is usually assumed that at least one message that hashes to the given value exists as shown in

Fig. 3. Therefore, to be resistance to pre-image attacks, one often says that the adversary (also called the attacker) is given $y = H(M)$ for some (randomly chosen) message M , which the attacker does not know. In other words, the attacker should find it is not possible to gain original data (or message (M)) from a given hash value $H(M)$.

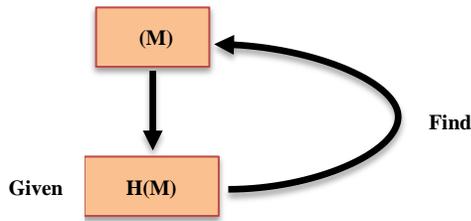


Fig. 3. Preimage resistance.

C. Resistance to Second Pre-Image Attacks

A second preimage is a message that hashes to the same value as a given (randomly chosen) message, called the first preimage. Obviously, the second preimage must be different from the first. Here, we assume that the attacker is also given the hash value of the first preimage. If not, then the attacker can compute it himself. In the latter case the cost of hashing the first preimage is placed on the attacker, which we do not assume here. A brute force preimage attack can also be used to find a second preimage.

One simply ignores the first preimage, except that one may take care not to try a message that is identical to the first preimage. By selecting messages at random, assuming that the domain of the hash function is much larger than the co-domain, the probability of the second preimage being equal to the first is negligible, and therefore we usually ignore this possibility. Due to the above attack, finding a second preimage seems to never be harder than finding a (first) preimage. However, there are artificial constructions that allow preimages to be found in constant time, but which are collisions and second preimage resistant.

Fig. 4 shows that the hash value $H(M)$ could change with the slightest change in message (M). In summary, it should be impossible for an attacker—which has been given a message to obtain the original digest after manipulating it.

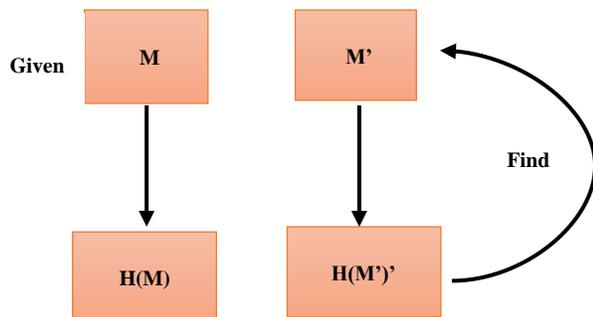


Fig. 4. Second preimage resistance.

Besides these properties, the hash function should also be able to work and calculate the digest for any input message of any size; the hash calculation process must be efficient.

III. APPLICATIONS OF HASH FUNCTION

Hash functions are used in many applications such as digital signature, message integrity, and authentication. This section discussed these applications.

A. Digital Signature

This is the first application of a secure hash function, it is a mathematical scheme used to validate the authenticity of the sender, message and signer of the document identity. In cases where it is crucial that an altered document or message is detected, or in any financial transaction, digital signature is commonly implemented. The signature for a document is produced via public and private keys utilized by the Digital Signature. This indicates that without authorization, it is difficult for another person to duplicate the document or message created by the person who had signed it first [7].

B. Message Integrity

Integrity checking is the foremost and fundamental objective of the hash function, which allows the detection of any changes being made to the data. The integrity of a message that is transmitted is checked via the sender, who hashes the message, whereby both hash value and message are sent. The message is generally sent from an insecure line, and only sometimes from a secure one. The received message is hashed from the side of the receiver, who checks the received hash value against the resulting hash value. The preservation of the message depends on whether or not the two hash values match; a match indicates preservation while a mismatch indicates non-preservation. There is a very low possibility that hash value and message are both altered (the hash value of the altered message is the altered hash value).

C. Message Authentication Code (MAC)

In constructions involving the Message Authentication Code (MAC), hash functions are popularly used as building blocks. Verification of identical sent and received messages can be done using the Message Authentication Code. However, only the sender or the recipient can compute the MAC. Therefore, identity verification of a sender to a third party cannot be executed using MAC. A keyed-hash function (which includes a keyed-in addition input to the message) is used to compute the MAC. The key must be kept secret or the operation will fail. Although third parties will not be privy to this key (kept secret from them), the same key must be used from the recipient and sender side. In the process of generating a MAC, any applicable constant string and hash function is used for the sender to input message and key. This is followed with the sending of the message and generated MAC to the receiver. The same hash function and algorithm is used on behalf of the receiver to generate a MAC of the message, so as to eliminate any chances of the message being manipulated. The message indeed would not have been manipulated if the MAC received from the sender matches the MAC generated by the receiver. This provides a simple way for verifying message integrity. To ensure MAC computation efficiency from both, the sender and receiver side, an efficient and high-speed hash function is required [6].

IV. STANDARD HASH FUNCTION ALGORITHMS

The most standard hash functions used today are the dedicated hash, that is, hash functions that are especially designed for hashing purpose only. In this section, we will describe the more popular hash functions.

A. MD5(Message Digest 5)

MD5 is a popular hash function in the MD family, designed by Rives in 1991. This hash function uses the Merkle–Damgård construction. The MD5 algorithm outputs a 128-bit length from an input of an arbitrary length message. However, several attacks have been found on MD5. In 1992, Bore and Bosselaers found collision attacks usually targeting the compression function. In 1996, Dobbertin published the fact that collision attacks targeted MD5. Successful collision attacks were also reported against MD5 in [8]. The improvement of collision attacks on MD5 were also found in previous works [9], [10].

B. RIPEMD-160

It is a well-known hash function in the RIPEMD family, designed by Dobbertin, Bosselaers and Perneel in 1996. It is part of the international standard ISO/IEC10118-3:2004 of dedicated hash functions. It also uses the Merkle–Damgård construction. It produces a message digest length of 160 bits [11]. However, semifree-start collision, preimage and collision attacks on RIPEMD-160 were found in [10].

C. Secure Hash Algorithm (SHA)

Secure Hash Algorithm (SHA) is a group of hash functions published by the National Institute of Standards and Technology as a US Federal Information Processing Standard (FIPS). All of the current SHA algorithms were developed by the NSA:

- **SHA-1:** NIST (1995) developed the Secure Hash Algorithm 1 or SHA-1, which also uses the Merkle–Damgård construction as MD5, and generates a 160-bit message digest for an arbitrary length input message.

However, collision attack was also founded against SHA-1 in previous studies [12]-[14]. Therefore, NIST announced the step-by-step elimination of SHA-1 [15].

- **SHA-2:** NIST (2002) added other algorithms to the SHA family with respective hash code lengths of 256, 384, and 512 bits i.e. SHA-256, SHA-384 and SHA-512, respectively. These follow the same structure as MD5 and SHA-1, but are more complex since a nonlinear function is added to the compression function. However, SHA-2 is not preferred to ensure integrity, as it is not as time efficient as SHA-1 [16]. On the other hand, Bitcoin, as the most popular cryptocurrency, uses SHA256 for Hashcash which provides security over transactions made between peers in the Bitcoin network. However, SHA-256 has no multi-threading ability, and thus it is not fast enough for transactions [17]. The most recent attacks on SHA-2 have been shown in previous works [18].
- **SHA-3:** After several successful collision attacks which were progressively reduced in complexity (such as MD5, SHA-1 and SHA-2), NIST, in the Federal Register, announced a public competition to develop SHA-3, a completely new hashing algorithm. In 2007, the announcement for the initiative was published. Then, four years later, on October 2nd, 2012, the winner of the competition Keccak, was announced. In 2014, NIST considered SHA-3 as a standard hash function. However, this algorithm is susceptible to first collision-finding attacks [19], [20]. On the other hand, the algorithm shows relatively low software performance compared to other hash functions [21].

V. COMPARISON OF STANDARD HASH ALGORITHM

The comparison of the standard hash algorithm based on the general properties, including block size, word size, output size, logical operation, and the number of rounds as shown in Table I. And also common attacks on these algorithms are summarized as illustrated in Table II.

TABLE I. COMPARISON BETWEEN STANDARD HASH FUNCTION ALGORITHMS BASED ON PROPERTIES

Properties	Name of Algorithm				
	MD5	RIPEMD -160	SHA-1	SHA-2 256/512	SHA-3 256/512
Block Size	512 bits	512 bits	512 bits	512/1024 bits	1088/576 bits
Word Size	32 bits	32bits	32bits	32/64 bits	320/320bits
Output Size	128bits	160 bits	160 bits	256/512 bits	1600/1600bits
Rounds	18	80	80	64/80	24/24
Operations	ADD,XOR, AND,OR, NOT, SHIFT	ADD,, ROTATE, XOR,AND, OR,NOT	ADD, XOR AND, OR,NOT, ROTATE.	ADD, XOR, OR, AND SHIFT,, ROTATE	-
Construction	Merkle-Damgard	Merkle-Damgard	Merkle-Damgard	Merkle-Damgard	Sponge

TABLE II. COMMON ATTACKS ON STANDARD HASH FUNCTION ALGORITHMS

Algorithm		Type of attacks	Complexity	References
MD5		Collision	2^{39}	Ref [8]
		Fast Collision	2^{18}	Ref [9]
RIPEM-160		Collision	2^{67}	Ref [22]
		Preimage	$2^{158.91}$	Ref [23]
SHA-1		Collision	$< 2^{69}$	Ref [12]
		Collision	2^{61}	Ref [13]
		Freestart Collision	-	Ref [14]
SHA-2	256	Preimage	$2^{255.5}$	Ref [24]
	512	Preimage	$2^{511.2}$	Ref [24]
SHA-3	256	Practical Collision and near-Collision	-	Ref [19]
	512	Possibility first Collision	-	Ref [20]

From the above discussion, it is found that most of the popular hash functions from different families suffer from collision attacks and also are not time efficient. As a solution to this problem, researchers proposed other algorithms.

VI. DISCUSSIONS

Many researchers have proposed their own algorithm in order to overcome the above issue as shown in Table III. In this section, the authors have discussed some of the variations in hash function algorithms.

Belfedhal and Faraoun [25] used a variant of the Merkle-Damgard construction basing off on cellular automata to introduce a hash function algorithm producing a 256-bit hash value. Although the algorithm yielded good results for statistical test, it was not tested against collision and preimage attacks.

Li et al. [26] used a dynamic S-box to design a chaotic hash function that produces 128-bit hash values as the final hash code and thus compromising its practicability and

flexibility lent via the S-box. One major drawback of this proposed algorithm is that the length of the hash code is not enough to guarantee security against collision or second pre-image attacks.

Abdulah et al. [27] developed a new hash function based on MD5, generating a 224-bit hash value. Perhaps the most serious disadvantage of this development is the time required to produce the message digest, which is as much as the MD5, meaning that the efficiency is very low.

Tur and Javurek [28] used neural network to develop hash function generation, which produced a 128-bit hash value. However, approaches of this kind are very difficult to execute besides having a short hash value.

Ahmad et al. [29] had integrated 2D and 1D chaotic maps in the development of a novel hash function scheme, where 128-bit hash value for an arbitrary length message was generated. Nevertheless, the length of the hash value is short and thus it is not resistant against collision attacks.

TABLE III. COMPARISON OF VARIATIONS OF HASH FUNCTION ALGORITHMS

Author	Year of publication	Advantages	Limitations
Belfedhal and Faraoun [25]	2015	It provides good cryptographic properties such as pseudo-random behavior and sensitivity to the input changes.	It was not tested against attacks that are cryptographic in nature e.g., meet-in-the-middle attacks, collision or birthday attacks.
Wang et al. [31]	2015	It provides variable output.	It was not tested against common attacks such as collision.
Li et al. [26]	2016	It has good statistical performance and collision resistance.	Any attacker could launch exhaustive collision attacks on the function because the final hash value is 128 bits
Tur and Javurek [28]	2016		Extra modules are still required to enable the proposed system to be used as a real application. It was not tested against attacks that are cryptographic in nature e.g., meet-in-the-middle attacks, collision or birthday
Li and Liu [30]	2016	The confusion and diffusion property of the proposed algorithm hash is good.	Any attacker could launch exhaustive collision attacks on the function because the final hash value is 128 bits.
Ahmad et al. [29]	2017	It has great statistical performance. It can generate hash value of length 160,256 or 512 bits.	
Zhang et al.[2]	2017	The proposed algorithm satisfies the requirement of statistical performance.	The algorithm is not time efficient to obtain the hash value.

Li and Liu [30] used chaotic mapping that is generalized and parameters that are variable to propose a hash function. In their work, an arbitrary length message was converted to corresponding ASCII values in the process of executing 6-unit iterations that has variable message values and parameters. Towards achieving the final hash value, iteration state values were used to cascade extracted bits. Based on a definition of the birthday attack, 128-bit hash value is not enough to guarantee a secure algorithm.

Wang et al. [31] proposed a hash algorithm, which generated a variable size digest. Their proposed algorithm was an improved version of MD-5 algorithm on the output length. Although the proposed algorithm has a variable size digest, which is good property to increase security, however it still uses the same construction as MD5.

Many researchers have proposed their own hash function algorithm as shown in Table III. Some of them were based on the chaotic design, with some being based on the complex chaotic system, the chaotic neural network, and the chaos tent map. Some of the current hash function designs produce a hash value size of 128 bits only which is not secure enough against collision attacks. Although these algorithms have offered satisfactory statistical performances, they are still weak in resisting collision attacks. Consequently, it is necessary to develop new approaches to hash function algorithm design that is able to prevent attacks effectively in comparison to existing algorithms as they are not sufficient to meet the requirement of latest technologies and security concern.

VII. CONCLUSION

There are various types of hash functions algorithms used to ensure the integrity and authentication of messages. Some have emerged as the standard, such as MD5, SHA-1, SHA-2 and SHA-3. This paper discusses these algorithms. It was found that most of them are either breakable, or are not time efficient. Also, this paper discusses other hash algorithms which were presented by researchers, but most of them were not tested against attacks that are cryptographic in nature such as collision attacks. Therefore, it can be concluded that a hash function that is efficient and safe, and fulfills application requirements such as data integrity and authenticity, must be designed and made into a priority.

ACKNOWLEDGMENT

We would like to express our gratitude to Ministry of Education (MOE Malaysia) for providing financial support (research grant Q.K130000.2538.19H12) in conducting our study. Our special thanks to Universiti Teknologi Malaysia (UTM) and specifically Advanced Informatics Department in Razak Faculty of Technology and Informatics for realizing and supporting this research work.

REFERENCES

[1] S. Mohammed, "Secure Hash Design & Implementation Based On Md5 & Sha-1 Using Merkle - Damgard Construction," *Int. J. Adv. Res. Comput. Sci. Technol. (IJARCST 2016) Vol.*, vol. 4, no. 2, pp. 92–94, 2016.

[2] P. Zhang, X. Zhang, and J. Yu, "A Parallel Hash Function with Variable Initial Values," *Wirel. Pers. Commun.*, vol. 96, no. 2, pp. 2289–2303, 2017.

[3] R. Haddaji, "Comparison of Digital Signature Algorithm and Authentication Schemes for H. 264 Compressed Video," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 9, pp. 357–363, 2016.

[4] R. P. Arya, U. Mishra, A. Bansa, and W. S. Email, "A Survey on Recent Cryptographic Hash Function Designs," *International J. Emerging Trends Technol. Comput. Sci.*, vol. 2, no. 1, pp. 1–6, 2013.

[5] N. Garg and N. Wadhwa, "Design of New Hash Algorithm with Integration of Key Based on the Review of Standard Hash Algorithms," *Int. J. Comput. Appl. (0975 – 8887)*, vol. 100, no. 8, pp. 11–18, 2014.

[6] M. Ghebleh and A. Kalso, "A structure-based chaotic hashing scheme," *Nonlinear Dyn.*, pp. 27–40, 2015.

[7] A. A. Alkandari, I. Al-shaikhli, and A. Alahmad, "Cryptographic Hash Function : A High Level View," *Int. Conf. Informatics Creat. Multimed. Cryptogr.*, pp. 129–135, 2013.

[8] X. Wang and H. Yu, "How to Break MD5 and Other Hash Functions," *EUROCRYPT 2005, LNCS 3494*, pp. 19–35, 2005.

[9] T. Xie, F. Liu, and D. Feng, "Fast Collision Attack on MD5," pp. 1–12, 2013.

[10] V. Chiriaco, A. Franzen, and R. Thayil, "Finding Partial Hash Collisions by Brute Force Parallel Programming," in *37th IEEE Sarnoff Symposium 2016, Newark, NJ, September 19-21, 2016*, vol. 5, pp. 1–6.

[11] F. Mendel, T. Peyrin, and M. Schl, "Improved Cryptanalysis of Reduced RIPEMD-160," *Int. Conf. Theory Appl. Cryptol. Inf. Security.*, pp. 484–503, 2013.

[12] X. Wang, Y. L. Yin, and H. Yu, "Finding Collisions in the Full SHA-1," *Int. Assoc. Cryptologic Res.* 2005, no. 90304009, pp. 17–36, 2005.

[13] M. Stevens, "New Collision Attacks on SHA-1 Based on Optimal Joint Local-Collision Analysis," *Int. Assoc. Cryptologic Res.* 2013, pp. 245–261, 2013.

[14] M. Stevens, P. Karpman, and T. Peyrin, "Freestart collision for full SHA-1," *EUROCRYPT 2016.*, vol. 2012, pp. 1–21, 2016.

[15] K. Wu, Y. Li, L. Chen, and Z. Wang, "Research of Integrity and Authentication in OPC UA Communication Using Whirlpool Hash Function," *Appl. Sci. ISSN2076-3417*, pp. 446–458, 2015.

[16] S. Verma and G. Prajapati, "Robustness and Security Enhancement of SHA with Modified Message Digest and Larger Bit Difference," in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016, pp. 0–4.

[17] G. Meliolla, K. A. Nugroho, and F. I. Hariadi, "Implementation of Hash Function on Embedded- System Platform using Chaotic Tent Map Algorithm," in *Electronics and Smart Devices (ISESD)*, *International Symposium on*, 2016, pp. 179–183.

[18] N. Kishore and B. Kappor, "Attacks on and Advances in Secure Hash Attacks on and Advances in Secure Hash Algorithms," *IAENG Int. J. Comput. Sci.*, no. September, 2016.

[19] I. Dinur, O. Dunkelman, and A. Shamir, "New attacks on Keccak-224 and Keccak-256," pp. 1–27, 2011.

[20] I. Dinur, O. Dunkelman, and A. Shamir, "Collision Attacks on Up to 5 Rounds of SHA-3 Using Generalized Internal Differentials," no. 827, pp. 1–30, 2013.

[21] D. Kim, D. H. B. J. Lee, and W. Kim, "LSH : A New Fast Secure Hash Function Family," *Springer Int. Publ. Switz.* 2015, pp. 286–313, 2015.

[22] F. Liu, F. Mendel, and G. Wang, "Collisions and Semi-Free-Start Collisions for Round-Reduced RIPEMD-160," in *23rd Annual International Conferences on Theory and Application of Cryptology and Information Security, ASIACRYPT2017, 2017*, pp. 1–32.

[23] Y. Shen and G. Wang, "Improved Preimage Attacks on RIPEMD-160 and HAS-160," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 2, pp. 727–746, 2018.

[24] D. Khovratovich, C. Rechberger, and A. Savelieva, "Bicliques for Preimages : Attacks on Skein-512 and the SHA-2 family," pp. 244–263, 2012.

- [25] A. E. Belfedhal and K. M. Faraoun, "Building Secure and Fast Cryptographic Hash Functions Using," *J. Comput. Inf. Technol.*, pp. 317–328, 2015.
- [26] Y. Li, G. Ge, and D. Xia, "Chaotic hash function based on the dynamic S-Box with variable parameters," *Nonlinear Dyn.*, vol. 84, no. 4, pp. 2387–2402, 2016.
- [27] H. S. Abdulah, M. A. H. Al-rawi, and D. N. Hammod, "Message Authentication Using New Hash Function," *J. Al-Nahrain Univ.*, vol. 19, no. 3, pp. 148–153, 2016.
- [28] M. Tur and M. Javurek, "Hash Function Generation by Neural Network," in *2016 New Trends in Signal Processing (NTSP)*, 2016.
- [29] M. Ahmad, S. Khurana, S. Singh, and H. D. Alsharari, "A Simple Secure Hash Function Scheme Using Multiple Chaotic Maps," *3D Res.*, vol. 8, no. 2, pp. 1–15, 2017.
- [30] Y. Li, X. Li, and X. Liu, "A fast and efficient hash function based on generalized chaotic mapping with variable parameters," *Neural Comput. Appl.*, vol. 28, no. 6, pp. 1405–1415, 2016.
- [31] M. Wang and Y. Zhen Li, "Hash Function with Variable Output Length," in *2015 International Conference on Network and Information Systems for Computers*, 2015, pp. 3–6.

An Analysis of Cloud Computing Adoption Framework for Iraqi e-Government

Ban Salman Shukur, Mohd Khanapi Abd Ghani, M.A. Burhanuddin

Faculty of Information Communication Technology
University Teknikal Malaysia Melaka,
Hang Tuah Jaya, Durian Tunggal Melaka, MALAYSIA

Abstract—This paper presents an analysis of the factors which could have possible affect over the adoption of cloud computing via the Iraqi e-government. A conceptual framework model for cloud computing within Iraqi e-government is proposed, analyzed, evaluated and discussed.

Keywords—e-Government; cloud computing; framework; Iraq; Iraqi e-government

I. INTRODUCTION

The government of Iraq realizes the significance of the e-government concept and the role of e-government to serve the Iraqi citizens and started the e-government project in 2004 after signing a memorandum between the Iraqi Ministry of Science and Technology (IMST) and the Italian government [1].

Unfortunately, like many growing countries, Iraqi's e-government system still has a lot of obstacles, problems, challenges and difficulties which affect its development [2], [3].

Poor management of resources, infrastructure problems, lack in IT specialists who are required for developing and maintaining the Iraqi e-government system beside the problem of growing amount of data are some examples of obstacles which the Iraqi e-government project is actually facing [2]-[6].

Cloud computing, as a new technology has changed the way that organizations think about and use ICT from many perspectives. Within this subject, e-governments around the world are really looking into cloud computing as a solution for many problems like increasing efficiency, reducing cost, providing more reliable and efficient services, and reducing cycle time [7].

This paper presents a review of the factors which could have possible affect over the adoption of cloud computing via the Iraqi e-government in a new point of view to overcome its obstacles which already exist. A conceptual framework of cloud computing adoption for Iraqi e-government is proposed and evaluated by IT specialists in IMST who are actually in charge of this project in Iraq.

II. E-GOVERNMENT

A. E-Government Systems

E-government systems have rapidly been implemented by developed as well as developing countries in different ways. Still, these implementation programs have varied widely in

their scope, effectiveness, levels and benefits which they provide [8].

Shareef et al. defined e-government as a way for governments to use the most innovative information and communication technologies, especially web-based internet applications to provide their citizens and businesses with more suitable access to government information and services [9]. While the Organization for Economic Co-operation and Development (OECD) uses the term "e-government" to the internet usage in making government's services and/or information available for its citizens[10].

Most developing countries have experienced problems in implementing e-government systems, and most have remained at the early stages of their implementation [11]. Krishnan and Teo give a percentage for e-government systems' implementation in developing countries as a 35% for complete failure, 50% for partly failure and they consider that only 15% of such systems were considered to be successful [12].

Al-Khouri stated that the majority of e-government in Arab countries have failed and are stuck in the access phase of Forrester's maturity model, while the other evolved Arab countries in e-government are still in the early steps of the interaction phase [13]. Forrester's maturity model portrays three phases to evaluate e-government transformation era namely access, interaction, and integration.

B. Benefits of E-Governments

The adoption and use of the e-government strategy can provide significant benefits for government in delivering a more effective and efficient information and services to all e-government sectors. It enables government agencies to align their efforts as needed to improve service and reduce operating costs [14].

Some of possible benefits of e-government implementation could be summarized in Table I.

Cost effectiveness is at the top of benefits which could be gained from a successful e-government system. Financial benefits are gained through cost reduction in reducing government transaction operations, reducing citizens' wasted time and efforts. Government's agencies will perform enhanced service quality for citizens and business and minimizing corruption and get the benefit of sharing information between them. This will built a better trust between government and its citizens.

TABLE I. POSSIBLE BENEFITS OF E-GOVERNMENT IMPLEMENTATION

Benefits	Authors
Financial	[15], [17], [18]
Social	[15], [17], [18], [19]
Government agencies	[15], [16], [19], [20]
Economy	[17]
Time Reduction	[18], [19], [20]
Achieve government's objectives	[15], [16], [19]
Sharing information	[16]
Built trust between government & citizens	[16]
Improve efficiency	[16], [19], [20]

C. Barriers of E-Government Implementation

Implementing any e-system could also have its own barriers and challenges. The implementation, development and management of an e-government system could have possible barriers and challenges which could be classified to major barriers. Each of these barriers has its own sub-barriers and all of them must be taken into great attention. Technical barriers are considered the most challengeable ones when implementing a new e-government system, they affect over the development and implementation of an effective e-government system, especially in a developing country like Iraq. Within this category we can count the readiness of ICT infrastructure, privacy and security as the most challengeable barriers which most literature reviews agreed on.

Other main category is the organizational category counted the top management support and the existence of qualified people to develop, manage, supervise and maintain the e-government system as sub-category's barriers do affect the implementation of e-government system, especially in developing and Arab countries like Iraq.

Table II shows a summary of possible barriers and challenges of E-Government implementation.

TABLE II. POSSIBLE BARRIERS OF E-GOVERNMENT IMPLEMENTATION

Categories	E-government Barriers	Authors
Technical	The Creation of ICT Infrastructure	[20], [21]
	Privacy	[22], [23]
	Security	[14], [24], [22], [20], [25], [26]
Organization	Top management support	[2], [27], [19], [29]
	Lack of Qualified Personnel and Training	[28], [29], [30]
	Lack of Collaboration	[28], [31], [32]
	Resistance to change to electronic ways	[28], [30]
Social	Digital Divide	[19], [33]
	Culture	[19], [34]
Financial Barriers	Cost	[35], [36]
Regulation and Policy Issues	Laws and legal subjects	[37]

III. CLOUD COMPUTING BENEFITS AND CHALLENGES FOR E-GOVERNMENT

The most official and countable definition about cloud computing was introduced by the National Institute of Standard and Technology (NIST), which defined it as

“Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction” [38].

For an effective e-government system deployment, some necessary facilities and characteristics are vital to be available; such as easy maintenance, reliability, mobile access, and more available data storage. Beside that for most of developing countries, the cost of construction an e-government system is considered a major difficulty against establishment.

Cloud computing can offer great benefits for e-government systems and solve problems of high cost establishment, ease of use, more data storage, mobile access, scalability and automatic backup and recovery [39]-[43].

For that, researchers, IT specialists and other involved stakeholders evolved many researches and studies concerning the adoption of cloud computing via e-government systems.

Today, many governments in the world are already using cloud services or adopting some kind of cloud computing solutions within their e-governments' systems. Other governments are investigating and studying the possible affecting factors about adopting cloud computing within their e-governments' systems through strategic migration plans to assist their required daily functions and improve the service delivery level for their citizens [44].

Accordingly, the affecting factors about this adoption should be clearly discovered, discussed and understood before taking the adoption decision. A strategic mechanism should be held and a migration strategy into the new cloud computing adoption technology must be completely investigated in order to welcome a good decision making when stepping towards cloud computing adoption in an e-government system.

Possible benefits and positive influences in adopting cloud computing via an e-government system is shown in Table III.

TABLE III. BENEFITS IN ADOPTING CLOUD COMPUTING VIA E-GOVERNMENT SYSTEM

Benefits	Authors
Easy to use & easy Implementation	[46], [47], [48], [49], [42], [53]
Scalability	[45], [46], [47], [48], [51], [42], [53]
Cost Reduction & Saving	[45], [46], [47], [48], [49], [42], [53]
Backup and Recovery	[45], [47]
Reduce local Data storage	[45], [51], [54], [50]
Availability & Accessibility	[45], [46], [47], [48], [42], [53]
Green computing	[45], [46], [47], [49], [52]

TABLE IV. POSSIBLE CHALLENGES IN ADOPTING CLOUD COMPUTING VIA E-GOVERNMENT SYSTEM

Challenges and Barriers	Authors
Security	[45], [48], [42], [54]
privacy	[45], [48], [42], [54]
Internet dependency	[55], [64]
Open standards and interoperability	[45], [42]
Law and National and International Regulatory	[54]
Business Reliable issues	[48], [42], [50], [54]
Identity and access management	[45]

Although the attention of most countries in adopting cloud computing within e-government systems is significantly rising, still there are some countries which have a little knowledge and leadership's concerns regarding this adoption [55].

Most of governments' concerns come from the fact that they are uploading their confidential information to their e-government systems and by adopting cloud computing within e-government system, the process of this information could be somewhere else outside the country beside the possibility of having a third party access to information.

This leads us to specify that security and privacy are the most important challenges which any government concern about when they are willing to adopt or use cloud within their e-government systems [56], [57].

Internet dependence and regularity issues are other concerning matters which should be encountered before choosing the appropriate cloud service provider(s), cloud service model(s) and cloud service deployment model(s).

Possible challenges in adopting cloud computing via e-government system in Table IV.

IV. ADOPTION MODELS FOR CLOUD COMPUTING IN E-GOVERNMENT SYSTEMS

To predict the adoption of new technology, like cloud computing, many well-known theories in information system (IS) field could be suggested and/or applied. Choosing suitable technologies and methods are compulsory for a good adoption practice to gain its valuable results.

Iraqi e-governments, as mentioned before, has some limitations, obstacles, challenges and some problems regarding its development and implementation process, which obviously have affected negatively on its completeness [3], [5] and [2]. For that, adopting cloud computing in the Iraqi e-government system requires a comprehensive analysis regarding the factors that could influence such adoption.

Pointing out such factors, will eventually leads to propose a realistic valuable framework that could be used to assist Iraqi government to step over problems and difficulties which the e-government system in Iraq is suffering from. Within this perspective, there are two adoption models are discussed in this paper and considered to be used for the proposition of Iraqi cloud computing e-government adoption framework.

TABLE V. SUGGESTED ADOPTION MODELS FOR CLOUD COMPUTING IN E-GOVERNMENT SYSTEM

Adoption Model of Cloud Computing in E-Government	Author
TOE framework +People	[44]
TAM+ (TRA)	[60]
TOE framework	[61]
TOE model + conceptual model of Cloud Computing adoption	[62]
TOE framework	[63]
Technological & Organizational factors	[64]
TAM	[65]

First, we discuss the Technology, Organization, and Environment (TOE) Framework, which is a theoretical framework that was developed by Tornatzky and Fleischer in 1990 and still widely used by IS researchers till now[58]. Within this framework, the components of an organization which is studying the adoption of new technology are classified into three major classifications; namely: Technological, Organizational and Environmental elements and the second model to be discussed here is the Technology Acceptance Model (TAM), which is mainly used to examine the user's interaction with the new innovation from two major points. The perceived usefulness (PU) which points to the degree of user's trust towards the new system and perceived ease of use (PEOU) which is the degree to which the user believes that using a particular system would be free of effort [59]. Table V shows a summary of suggested models in adopting cloud computing within e-government.

V. FACTORS AFFECTING THE ADOPTION OF CLOUD COMPUTING VIA E-GOVERNMENT

Scientific analysis with reliable bases should be implemented in order to drive a conceptual framework for the adoption of new technology. This matter is a big matter of importance, especially when the adoption decision could affect and be affected by a system which is employed and operated by a government and used by all citizens. In addition, the local, region, religion, culture and legal characteristics and issues beside the uniqueness of each country should be considered and investigated.

Because of all the mentioned points, studying factors which may have affects over the adoption of a new technology from different perspectives, as researchers went through should be considered and taken care for.

Factors have their own positive or negative impact over the adoption of any new technology. The adoption of cloud computing via e-government is not an exception of this.

Possible affecting factors which could affect the adoption of cloud computing via e-government, from literature is summarized in Table VI.

TABLE VI. FACTORS AFFECTING THE ADOPTION OF CLOUD COMPUTING VIA E-GOVERNMENT

Affecting Factors	Authors	
Technological context	Cost saving (+)	[66], [67], [70], [71], [73], [74], [75], [76], [46], [45], [77], [78], [69], [79], [47], [49], [80], [81], [68], [82], [64], [72], [83], [87]
	Scalability (+)	[66], [70], [71], [73], [75], [76], [46], [42], [74], [62], [79], [47], [80], [81], [61], [63], [82], [64], [72], [87]
	flexibility (+)	[67], [49], [50], [73], [74], [75], [62], [79], [80], [81], [44], [61], [63], [82], [64], [72], [83], [87]
	Compatibility(+)	[75], [78], [79], [81], [81], [61], [64], [72]
	Complexity(-)	[75], [78], [79], [80], [81], [61], [64], [72], [86]
	Security & privacy(-)	[66], [67], [70], [71], [73], [74], [75], [76], [42], [77], [78], [62], [44], [82], [63], [64], [86], [87]
	Resource Utilization (+)	[66], [67], [70], [73], [45], [77], [44], [70], [82]
Environmental Context	Reliable(+)	[67], [73], [74], [69], [80], [81], [61], [83], [86]
	Available (+)	[67], [73], [62], [79], [80], [81], [68], [83], [86]
	Ownership (-)	[67], [73], [74], [79], [81], [63]
	Mobile Access(+)	[66], [67], [71], [73], [76], [45], [62], [79], [46], [82]
	Migration(-)	[70], [71], [73]
	Internet Connection	[70], [82]
Organizational Context	Top Management (+)	[75], [78], [79], [80], [81], [44], [61], [70], [64], [72]
	IT infrastructure(+)	[75], [42], [80], [81], [61], [63], [64], [72]
	IT Human Resources(+)	[75], [70], [81], [44], [61], [63], [82]
Easy of Use (+)	[67], [70], [73], [74], [46], [42], [47], [49], [82], [83]	
Regulation Issues	[62], [70], [83]	

VI. FACTORS AFFECTING CLOUD COMPUTING ADOPTION VIA IRAQI E-GOVERNMENT

According to Table VI, there are 19 affecting factors which most researchers agreed on and they are divided into 5 major categories: namely, technological context, environmental context, organizational context, ease of use and regulation issues.

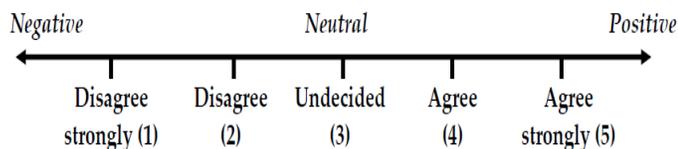


Fig. 1. Straightforward Notion for Likert [84].

However, these factors should be tested and evaluated by Iraqi IT specialists who are responsible of the e-government system within IMST to see if they are appropriate for Iraq or not. For that a questioner was conducted to evaluate these

factors by above mentioned employees besides semi structured interviews to reach common understandings and solutions.

The sample included 25 participants, 15 participants were male and 10 of them female who have different job titles; namely: programmers, IT engineers, managers and experts, all of them working within the Iraqi e-government project in IMST / Information Technology section. They work as technicians, technical administrations, information engineers, programmers, managers and experts and varied in age between 20-60 years old.

Likert scale, which is one of the most important and frequently used tools by researchers in sciences was used to test whether the factors that affect the adoption of cloud computing within the Iraqi e-government have positive or negative impact over the adoption [85].

Likert suggested that a successful attitude measurement is to carry out the fundamentals measurements to review the respondents' options which best reflect their position on the

measurement. Fig. 1 shows the Likert's measurements' notation [84].

For this paper's requirements, the respondents were asked to disclose the factors that could influence the adoption of cloud computing within e-government in Iraq. A 5-point Likert scale was used for questioner, where 1 point is accorded to strongly disagree, 2 points for disagree, 3 points for neutral (neither agree nor disagree), 4 points for agree and 5 points is a scale of strongly agree.

Table VII shows the main and sub categories with a suitable question statement representing the possible affecting factors which could have impact over the adoption of cloud computing via Iraqi e-government. This was the sample which

was delivered for IMST employers who were the participants in this survey accompanied by Likert 5-point scale to select the suitable selection by them.

The results from this part of questioner were analyzed according to Likert's scale and the mean as well as the standard deviation (S.D) were calculated to validate suggested factors. These results which are illustrated in Table VIII shows clearly that most of suggested factors (all except the migration factor) have an exceeded mean value (>3.0) which means they are acceptable by the IT experts and employees who are actually working in the Iraqi e-government project at IMST.

S.D also pointed out a non-acceptable value (< 0.7) for migration factor.

TABLE VII. SAMPLE OF QUESTIONER FORM

Technological Context	
Statement	
1	Is adopting cloud computing in Iraqi E-government system has a cost saving advantage impact factor over the system?
2	Is adopting cloud computing in Iraqi E-government system has a flexibility advantage impact factor over the system?
3	Is adopting cloud computing in Iraqi E-government system has a scalability advantages impact factor over the system?
4	Is adopting cloud computing in Iraqi E-government system has a compatibility advantages impact factor over the system?
5	Is adopting cloud computing in Iraqi E-government system has a complexity disadvantage impact factor over the system?
6	Is adopting cloud computing in Iraqi E-government system has a security disadvantage impact factor over the system?
7	Is adopting cloud computing in Iraqi E-government system has a privacy disadvantage impact factor over the system?
8	Is adopting cloud computing in Iraqi E-government system has a resource utilization advantage impact factor over the system?
Organizational Context	
Statement	
9	Is the top Management Support has a positive impact over the adoption of cloud computing within Iraqi E-government?
10	Is adopting cloud computing in Iraqi E-government system has a positive impact over the IT infrastructure?
11	Is adopting cloud computing in Iraqi E-government system has a positive impact over the IT Human Resources?
Environmental Context	
Statement	
12	Is adopting cloud computing in Iraqi E-government system has a Reliability advantage impact over the system?
13	Is adopting cloud computing in Iraqi E-government system has the availability advantage impact over the system?
14	Is adopting cloud computing in Iraqi E-government system has an Ownership disadvantage impact over the system?
15	Is adopting cloud computing in Iraqi E-government system has a Mobile Access advantage impact over the system?.
16	Is adopting cloud computing in Iraqi E-government system has a Migration disadvantage impact over the system?
17	Is adopting cloud computing in Iraqi E-government system can have the Internet Connection advantage impact over the system?
Ease of Use	
Statement	
18	Is adopting cloud computing in Iraqi E-government system has an Easy of Use advantage impact over the system?
Legal Issues	
Statement	
19	Is adopting cloud computing in Iraqi E-government system has a Legal Issue disadvantage impact over the system?

TABLE VIII. QUESTIONER RESULTS

Technological Context			
Statement		Mean	S.D
1	Is adopting cloud computing in Iraqi E-government system has a cost saving advantage impact factor over the system?	4.6	0.63245553
2	Is adopting cloud computing in Iraqi E-government system has a flexibility advantage impact factor over the system?	3.96	0.91564185
3	Is adopting cloud computing in Iraqi E-government system has a scalability advantages impact factor over the system?	4.52	0.75471849
4	Is adopting cloud computing in Iraqi E-government system has a compatibility advantages impact factor over the system?	4.04	1.28
5	Is adopting cloud computing in Iraqi E-government system has a complexity disadvantage impact factor over the system?	4.32	0.96829747
6	Is adopting cloud computing in Iraqi E-government system has a security disadvantage impact factor over the system?	3.64	1.05375519
7	Is adopting cloud computing in Iraqi E-government system has a privacy disadvantage impact factor over the system?	4.4	1.0198039
8	Is adopting cloud computing in Iraqi E-government system has a resource utilization advantage impact factor over the system?	4.6	0.48989795
Organizational Context			
Statement		Mean	S.D
9	Is the top Management Support has a positive impact over the adoption of cloud computing within Iraqi E-government?	3.64	1.29243955
10	Is adopting cloud computing in Iraqi E-government system has a positive impact over the IT infrastructure?	4.92	0.2712932
11	Is adopting cloud computing in Iraqi E-government system has a positive impact over the IT Human Resources?	4.44	0.89799777
Environmental Context			
Statement		Mean	S.D
12	Is adopting cloud computing in Iraqi E-government system has a Reliability advantage impact over the system?	3.52	1.38910043
13	Is adopting cloud computing in Iraqi E-government system has the availability advantage impact over the system?	4.2	1.16619038
14	Is adopting cloud computing in Iraqi E-government system has an Ownership disadvantage impact over the system?	4.08	1.16344317
15	Is adopting cloud computing in Iraqi E-government system has a Mobile Access advantage impact over the system?	4.4	0.8
Statement		Mean	S.D
16	Is adopting cloud computing in Iraqi E-government system has a Migration disadvantage impact over the system?	2.92	0.56
17	Is adopting cloud computing in Iraqi E-government system can have the Internet Connection advantage impact over the system?	3.56	0.89799777
Ease of Use			
Statement		Mean	S.D
18	Is adopting cloud computing in Iraqi E-government system has an Easy of Use advantage impact over the system?	3.92	0.97652445
Legal Issues			
Statement		Mean	S.D
19	Is adopting cloud computing in Iraqi E-government system has a Legal Issue disadvantage impact over the system?	3.64	0.97488461

REFERENCES

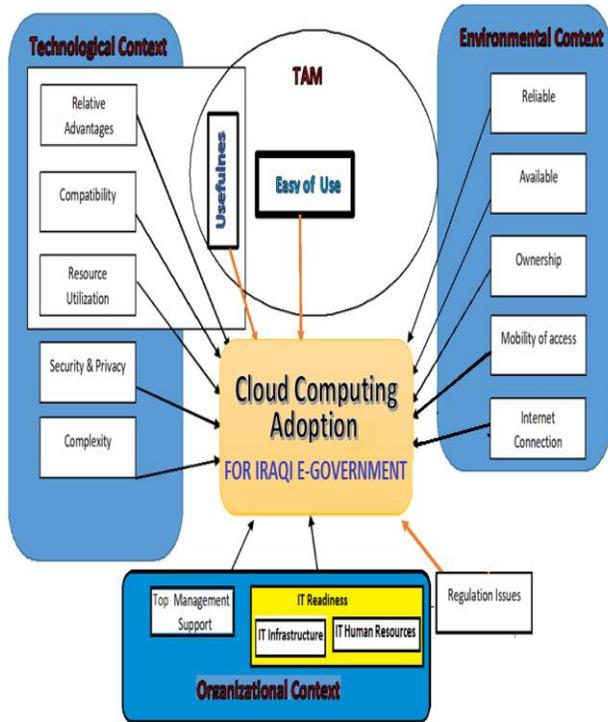


Fig. 2. Conceptual model for factors affecting cloud computing adoption for Iraqi e-government.

According to that, the migration factor was excluded from the conceptual framework model relating to the adoption of cloud computing in Iraqi e-government. Fig. 2 shows the resultant conceptual model for the factors affecting cloud computing adoption for Iraqi E-government.

VII. CONCLUSION

There are many factors which need to be considered before adopting cloud computing in a new environment.

For a developing country like Iraq, which has its own circumstances and issues, developing a conceptual framework about the adoption of cloud computing should rely own solid scientific bases and goes through a lot of investigation and validation by specialized IT expertise and employees.

The conceptual model for factors affecting cloud computing adoption in Iraqi e-government was derived upon deep and wide verity of related literature review and has its own bases and theories of adoption which should be accounted to assure the right adoption context for a new innovation technology. From analyzing factors we saw that a TOE framework and TAM model with addition and modifications which researchers did concerning the uniqueness characteristics about Iraqi e-government was suitable as a theoretical framework and a model to be tested.

The outcomes was very corrigible and most factors extracted through literature review (all except one) were conferred to be affected factors if a cloud computing is to be chosen via Iraqi e-government system.

- [1] A.H. Abdul-Alrahman, "Human Resources Investment as an Introduction to improve the efficiency & activity of workers in E – Government" Journal: Journal of the planner and development, Issue: 24, 2011.
- [2] A. Al-Dabbagh, "Former Spokesman: Iraqi Decision Makers Are Unqualified and Incompetent", 2014.
- [3] A.A. Hassan, "Status of E-Government in Iraq and What the Challenges of Development and Implementation", International Journal of Science and Research (IJSR), Volume 5 Issue 5, 2015.
- [4] H. Sabah, "Designing and Implementation Iraqi E-Government Front Office Online System", Journal of Knowledge Management, Economics and Information Technology, Vol. IV, Issue 2, 2014.
- [5] H.H. Mahmoud, "E-Government in Iraq", Journal of Engineering and Development, Vol. 14, No. 4, 2010.
- [6] T. E. Ibrahim, "A Road Map to a Successful Application of E-Government in Iraq". Çankaya University, 2014.
- [7] A. Bisong, S. Syed & M. Rahman, "An Overview of the Security Concerns in Enterprise Cloud computing", CoRR, 3(April 2012), 30–45
- [8] M. Alsaif, "Factors affecting citizens' adoption of e-government moderated by socio-cultural values in Saudi Arabia", University of Birmingham, 2014.
- [9] M.A. Shareef, V. Kumar, U. Kumar, & Y.K. Dwivedi," E-Government Adoption Model (GAM): Differing service maturity levels" Government Information Quarterly, 28(1), 2011, 17- 35.
- [10] OECD. OECD E-Government Flagship Report "The E-Government Imperative," Public Management Committee, Paris: OECD, 2003.
- [11] U. Sivarajah, & Z. Irani, "Exploring The Application Of Web 2.0 In E-Government: A United Kingdom Context", tGov2012: Brunel University, University Kingdom, 2012.
- [12] S. Krishnan, & T.S. "Moderating effects of governance on information infrastructure and e-government development", Journal of the American Society for Information Science and Technology, 63(10), 2012.
- [13] A.M. Al-Khoury, "E-government in Arab countries: A 6-staged roadmap to develop the public sector", Journal of management and Strategy, 4(1), 80, 2013.
- [14] N. Nkwe, "E-government: challenges and opportunities in Botswana", International journal of humanities and social science, 2(17), 2012, 39-48.
- [15] B.C. Smith, "Good Governance and Development", New York: Palgrave Macmillan, 2007.
- [16] P. Directorate, T. D. Committee, P. G. In, I., & P. Service, "OCDE. Governance, Innovation in public service delivery", OCDE, For Official Use, 2011.
- [17] E. Lau, "E-government and the drive for growth and equity", Proceedings Conference From E-Gov to I-Gov, organization for economic cooperation and development e-government project, 2015.
- [18] J. Lee, "10 year retrospect on stage models of e-Government: A qualitative meta-synthesis", Government Information Quarterly, 27(3), 2010, 220-230.
- [19] M. Alshehri, & S. Drew, "E-Government fundamentals". Paper presented at the IADIS International Conference ICT, Society and Human Beings, 2010.
- [20] T.M. Waema, & E.O. Adera, " Local governance and ICTs in Africa: case studies and guidelines for implementation and evaluation: IDRC", 2011.
- [21] H. Alsaghier, M. Ford, A. Nguyen, & R. Hexel, " Conceptualising citizen's trust in e-government: Application of Q methodology", Leading Issues in E-Government, 2011.
- [22] Y.K. Dwivedi, V. Weerakkody, & M. Janssen, " Moving towards maturity: challenges to successful e-government implementation and diffusion" ACM Sigmis Database, 42(4), 2012.
- [23] M.W. Donker-Kuijter, M. de Jong, & L. Lentz, " Usable guidelines for usable websites. An analysis of five e-government heuristics". Government Information Quarterly, 27(3), 2010, 254-263.

- [24] G. Udo, "Privacy Security Concerns as Major barriers for e-commerce: A survey study" in *Information Management & Computer Security*, vol.9, no.4, 2001.
- [25] U. Lofstedt, "E-government-assessment of current research and some proposals for future directions", *International journal of public information systems*, 1(1), 2012.
- [26] M. Mustafa Kamal, & M. Alsudairi, "Investigating the importance of factors influencing integration technologies adoption in local government authorities". *Transforming Government: People, Process and Policy*, 3(3), 2009, 302-331
- [27] R. Heeks, "Understanding and measuring e-government. International Benchmarking Studies, UNDESA workshop, E-Participation and E-Government: Understanding the Present and Creating the Future", Budapest, Hungary, 27-28 July 2006.
- [28] V. Ndou, "E-government for developing countries: opportunities and challenges", *The Electronic Journal on Information Systems in Developing Countries*; Volume 18, No.1, 2004.
- [29] A. Meijer, et al, "Government 2.0: Key Challenges to Its Realization. *Electronic Journal of e-Government*", 10(1), 2012, 59 – 69
- [30] S. Sharma, and J. Gupta, "Building Blocks of an E-government- A Framework", *Journal of Electronic Commerce in Organizations*, vol.1, no. 4, 2003.
- [31] S. Dorji, "E-government initiatives in Bhutan: Government to Citizen (G2C) service delivery initiative-A case study", Murdoch University, 2012.
- [32] A.N.H. Zaied, "Barriers to e-commerce adoption in Egyptian SMEs", *International Journal of Information Engineering and Electronic Business*, 4(3), 9, 2012.
- [33] F. Bélanger, & R.E. Crossler, "Privacy in the digital age: a review of information privacy research in information systems", *MIS quarterly*, 35(4), 2011, 1017-1042.
- [34] S. Jackson, "Organizational culture and information systems adoption: A three-perspective approach" in *Information and Organization*, 21(2), 57-83, 2011.
- [35] D.M. West, "State and Federal E-government in United States", 2001.
- [36] A. Carvin, J. Hill, and S.S. Smothers, "E-government for all: Ensuring equitable access to online government services", *The EDC center for media & community and the NYS forum*, 2004.
- [37] I.A. Alghamdi, R. Goodwin, & G. Rampersad, "E-government readiness assessment for government organizations in developing countries", *Computer and Information Science*, 4(3), 3, 2011.
- [38] P.Mell, & T. Grance, "The NIST definition of cloud computing", 2011.
- [39] KPMG, "The Cloud: Changing the Business Ecosystem", Kpmg, 1–102, 2011.
- [40] A. Abdulaziz, "Cloud computing for increased business value", *International Journal of Business and Social Science*, 3(1), 2012.
- [41] K. Vats, S. Sharma, & A. Rathee, "A Review of Cloud Computing and e-Governance", *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(2), 5, 2012.
- [42] J. Liang, "Government cloud: enhancing efficiency of e-government and providing better public services. Paper presented at the Service sciences (IJCSS)", international joint conference on Service Sciences, Service Innovation in Emerging Economy: Cross-Disciplinary and Cross-Cultural Perspective, 2012.
- [43] A. E. Youssef, "Exploring Cloud Computing Services and Applications", *Journal of Emerging Trends in Computing and Information Sciences*, 3, 2012.
- [44] R. Kurdi, A. Taleb-Bendiab, M. Randles & M. Taylor, "E-Government Information Systems and Cloud Computing (Readiness and Analysis)", In *Developments in E-Systems Engineering (DeSE)*, IEEE, 404–409, 2011.
- [45] T. Almarabeh, Y.K. Majdalawi, & H. Mohammad, "Cloud Computing of E- Government. *Communications and Network*", 8(8), 2016, pp1–8.
- [46] R.K. Das, S. Patnaik, & A.K. Misro, "Adoption of cloud computing in e-governance *Advanced Computing* (pp. 161-172): Springer, 2011.
- [47] R. Sharma, A. Sharma, & R.R. Singh, "E-Governance & Cloud Computing: Technology Oriented Government Policies", *IJRIM* 2(2), 2012.
- [48] S. Alshomrani, and S. Qamar, "Cloud Based E-Government: Benefits and Challenges", *International Journal of Multidisciplinary Sciences and Engineering*, 4(6): 2013, pp. 1-7.
- [49] V. Kundra, "State of Public Sector Cloud Computing", Federal Chief Information Officers Council, 2010.
- [50] A. Rastogi, "A model based approach to implement cloud computing in e-Governance", *International Journal of Computer Applications*, 9(7), 2010, pp15-18
- [51] K.L. Bansal, S.K. Sharma, and S. Sood, "Impact of Cloud Computing in Implementing Cost Effective E-governance Operations", *GIAN JYOTI E-JOURNAL*, 1(2), 2012.
- [52] M. Bellamy, "Adoption of Cloud Computing Services by Public Sector Organisations", Paper presented at the Services (SERVICES), 203 IEEE Ninth World Congress, 2013.
- [53] B. Zwattendorfer, & A. Tauber, "The public cloud for e-government", *International Journal of Distributed Systems and Technologies (IIDST)*, 4(4), 2013, pp1-14.
- [54] Al-ghanim, "Relationship and Cloud Factors Affecting Government Confidence in the Public Cloud", PhD Thesis, De Montfort University, United Kingdom, 2017.
- [55] N. ALMutairi, & S. Fahad Thuwaini, "Cloud Computing Uses for E-Government in the Middle East Region Opportunities and Challenges", *International Journal of Business and Management*, 10(4), 2015.
- [56] D. Woods, & P. Hofmann, "Cloud computing: the limits of public clouds for business applications", *Internet Computing, IEEE*, 14(6), 90-93, 2010.
- [57] S. Subashini, & V. Kavitha, "A survey on security issues in service delivery models of cloud computing", *Journal of Network and Computer Applications*, 2011.
- [58] L. G. Tornatzky, M. Fleischer, & A. K. Chakrabarti, "Processes of technological innovation: Lexington Books", 1990.
- [59] F.D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly* (13:3), 1989, pp. 319-339.
- [60] D.H. Shin, "User centric cloud service model in public sectors: policy implications of cloud services", *Government Information Quarterly*, 30(2), 2013, 194-203.
- [61] H. P. Borgman, B. Bahli, H. Heier, & F. Schewski, "Cloud rise: Exploring cloud computing adoption and governance with the TOE framework". In *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2013, pp. 4425–4435.
- [62] M. Alसानة, "Factors Affecting the Adoption of Cloud Computing in Saudi Arabia's Government Sector", 2015.
- [63] H. Trivedi, "Cloud computing adoption model for governments and large enterprises. Massachusetts Institute of Technology", 2013.
- [64] M. A. Wahsh, & J. S. Dhillon, "A systematic review of factors affecting the adoption of cloud computing for e-Government implementation", *ARPN Journal of Engineering and Applied Sciences*, 10(23), 2015a, 17824–17832
- [65] R. Alotaibi, L. Houghton, K. Sandhu, "Factors Influencing Users' Intentions to Use Mobile Government Applications in Saudi Arabia: TAM Applicability", *International Journal of Advanced Computer Science and Applications*. 8, 200–211, 2017.
- [66] K. Kavitha, "Study on Cloud Computing Model and its Benefits, Challenges", *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1), 2423–2431, 2014.
- [67] M. Oqail Ahmad, & R. Zaman Khan, "The Cloud Computing: A Systematic Review", *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*, 2015.
- [68] B. L. Sahu, & R. Tiwari, "A comprehensive study on Cloud computing", *International journal of Advanced Research in Computer science and Software engineering*, 2(9), 2012.
- [69] H.M. El-Bakri, "Cloud Computing in E-Government: A Survey", 2015

- [70] Shiny ,”Load Balancing In Cloud Computing:A Review , Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,2013
- [71] C. Ting, S. Xue, F. Tiong, & W. Xin, ” BENEFITS AND CHALLENGES OF THE ADOPTION OF CLOUD COMPUTING IN BUSINESS”, International Journal on Cloud Computing: Services and Architecture (IJCCSA), 6(6), 1–15, 2016.
- [72] M. Carroll, A. Van Der Merwe, &P. Kotzé, ” Secure Cloud Computing: Benefits, Risks and Controls”, Information Security for South Africa, 1–9, 2011
- [73] IDC, ” Cloud Computing 2010 an IDC Update.”, from IDC (International Data Corporation) Report. Update. IDC, 2009.
- [74] H. Gangwar, H. Date, and R. Ramaswamy,“ Understanding determinants of cloud computing adoption using an integrated TAM-TOE model”, Journal of Enterprise Information Management. Vol. 28, 2015, pp. 107-130
- [75] Y. Alshamaila, S. Papagiannidis, and F. Li, ” Cloud computing adoption by SMEs in the north east of England”, Journal of Enterprise Information Management [Online], 26(3), 2013, pp.250–275.
- [76] T. Almarabeh, Y.K. Majdalawi, & H. Mohammad, ” Cloud Computing of E- Government. Communications and Network”, 8(8), 1–8, 2016.
- [77] M. Alsanea, & D. Wainwright, ” Identifying the Determinations of Cloud Computing Adoption in a Government Sector- a Case Study of Saudi Organization”, International Journal of Business and Management Studies, Vol 6, No 2, 2014.
- [78] S.R. Tehrani, ” Factors Influencing the Adoption of Cloud Computing by Small and Medium-Sized Enterprises (SMEs)”, Ryerson University, 2013.
- [79] C. Low, Y. Chen, and M. Wu, ” Understanding the determinants of cloud computing adoption”, Industrial management & data systems. Vol. 111, 2011, pp. 1006-1023.
- [80] O. Harfoushi, A.H. Akhorshaidah, N. Aqqad, M.A. Janini, & R. Obiedat, ” Factors Affecting the Intention of Adopting Cloud Computing in Jordanian Hospitals”, Communications and Network, 08(02), 88–101, 2016.
- [81] F. Mohammed, and O. Ibrahim, “Models of adopting cloud computing in the e-government context: a review”, Jurnal Teknologi, 73 (2), 2015, pp. 51-59
- [82] H.M. El-Bakry, ”Cloud Computing in E-Government: A Survey”, 2015.
- [83] Mustafa Musa Jaber, Mohd Khanapi Abd Ghani1, Nanna Suryana Herman, “A Review Of Adoption Of Telemedicine In Middle East Countries: Toward Building Iraqi Telemedicine Framework, International Symposium on Research in Innovation and Sustainability”, (ISoRIS '14) 15-16 October 2014, Malacca, Malaysia, Sci.Int. (Lahore),26(5), 2014, pp. 1795-1800
- [84] R. Johns, “Likert Items and Scales. Survey Question Bank: Methods Fact Sheet 1”, 1–11,2010.
- [85] A. Joshi, S. Kale, S. Chandel, D. Pal, “Likert Scale Explored and Explained”, British Journal of Applied Science & Technology. 7, 396–403, 2015.
- [86] S. a. Alateyah, R. R. M. Crowder, G. B. G. Wills, “ An Exploratory study of proposed factors to Adopt e-government Services Saudi Arabia as a case study”, (IJACSA) International Journal of Advanced Computer Science and Applications. 2013, p. pp 57-66.
- [87] M. A. Alanezi, “Factors Influencing Cloud Computing Adoption in Saudi Arabia ’ s Private and Public Organizations : A Qualitative Evaluation”. (IJACSA) International Journal of Advanced Computer Science and Applications ,9, 121–129, 2018.

A Survey on Tor Encrypted Traffic Monitoring

Mohamad Amar Irsyad Mohd Aminuddin¹,
Zarul Fitri Zaaba*², Manmeet Kaur Mahinderjit Singh³
School of Computer Sciences
Universiti Sains Malaysia,
11800 USM, Pulau Pinang, Malaysia

Darshan Singh Mahinder Singh⁴
Centre for Drug Research
School of Computer Sciences
Universiti Sains Malaysia,
11800 USM, Pulau Pinang, Malaysia

Abstract—Tor (The Onion Router) is an anonymity tool that is widely used worldwide. Tor protect its user privacy against surveillance and censorship using strong encryption and obfuscation techniques which makes it extremely difficult to monitor and identify users' activity on the Tor network. It also implements strong defense to protect the users against traffic features extraction and website fingerprinting. However, the strong anonymity also became the heaven for criminal to avoid network tracing. Therefore, numerous of research has been performed on encrypted traffic analyzing and classification using machine learning techniques. This paper presents survey on existing approaches for classification of Tor and other encrypted traffic. There is preliminary discussion on machine learning approaches and Tor network. Next, there are comparison of the surveyed traffic classification and discussion on their classification properties.

Keywords—Encrypted traffic monitoring; Tor; machine learning; security; survey

I. INTRODUCTION

Tor (The Onion Router) is a well-known anonymity network globally [1]. The popularity of Tor is undeniable as the Tor's users increase up to two million just in year of 2017 alone [2] and currently there are more than four million users worldwide. Similar with other anonymity project such as I2P [3] and FreeNet [4], Tor primary goal is to provide users with privacy protection and anonymous access to the Internet. This will facilitate the Internet users with mechanism to hide their activity thus protecting their privacy to some extent, i.e. to hide the source, the destination, and the nature of the communication, other than encrypting the content itself [35]. Although there are lots of good usage practice, Tor however are also dual-use networks just like other technology such as BitTorrent (where users are not only use it to share free materials, they also share copyrighted materials) since it has been exploited for illegal activities purposes [5]-[8].

Due to the complexity nature of the Tor encrypted traffic, the research community has put considerable effort on analyzing the Tor security especially on the possibility of deanonymizing the Tor users [11]-[15]. These researches have focused on decoy traffic [12], exit router logging [11], attack on identifying Tor relays [14] and investigation on exit relays trustworthiness [15]. Although the result is promising, but these techniques lack of security monitoring proficiencies on the Tor network. In security monitoring perspective, the Tor traffic in general should be monitored and analyzed to obtain useful knowledge such as information on the website that

Tor's user access. Therefore, there are few has focused on the privacy disclosure based on identifying application information of traffic in the Tor network [10]. This approach would allow for the Tor traffic to be monitored in large scale and real-time. Although it is not directly deanonymized the user activity on the Tor network, learning the application information that being used by the users in the Tor network is a part of the Tor privacy concern [33]. This security monitoring capabilities could be achieved using machine learning classification technique similar to the classification of encrypted traffic on the surface web.

Even though machine learning classification for encrypted traffic has been studied intensively [17], [48]-[50], the process of applying these studied to classify the Tor traffic is remarkably challenging due to valuable traffic features that obtained from previous study are irrelevant in the context of Tor network and Tor itself developed with strong anonymity protection [31].

The main objective of this paper is to survey the application of machine learning techniques for encrypted Tor traffic classification and several latest clearnet encrypted traffic classification studies. Based on the results of the investigation, we will discuss and compared comprehensively on those machine learning techniques and its operation.

The rest of the paper is organized as follows. In Section 2, will be discussion on the technical background of traffic encryption on the Tor network. In Section 3, we discuss the fundamental of machine learning in traffic classification. Section 4 contain the discussion of studies on machine learning techniques. Section 5 contains the comprehensive comparison and consideration on these surveyed techniques. Finally, the paper is concluded in Section 6.

II. TOR BACKGROUND

A. Onion Routing

As Tor created to provide anonymity services that allow people to improve their privacy and security on the Internet, it is run by group of volunteer-operated servers around the world. The main backbone of Tor network is the distributed relay server (Tor node) which providing the onion routing capabilities. Onion routing is a concept of anonymous communication over a computer network where the messages are encapsulated in multiple layers of encryption [18]. The encrypted data is transmitted through series of Tor nodes (current Tor implementation use three nodes [1]) which is called as a Tor circuit. Each of the node will decrypt a layer of

encryption to uncover the next destination of the traffic without the knowledge on whether the source of the traffic is coming from a Tor client or from another Tor node. Only the exit (third) node know the true destination of the messages. Hence, there are no node that has both information on the source and destination of the messages.

Fig. 1 shows the Tor circuit example. Notice that the connection between the exit node and destination server does not encrypted. This is because Tor provide traffic encryption mechanism while the messages are in the Tor network [19]. The moment that messages go out of the Tor network, it is up to the users whether the traffic is encrypted or not. As an example, if a user accesses a HTTPS (Hyper Text Transfer Protocol Secure) websites, not only the communication is encrypted in the Tor network, but also encrypted outside of the Tor network through the HTTPS. If the user accesses an unsecure HTTP website, the communication is encrypted only in the Tor network and no encryption provided outside of the Tor network (situation in Fig. 1).

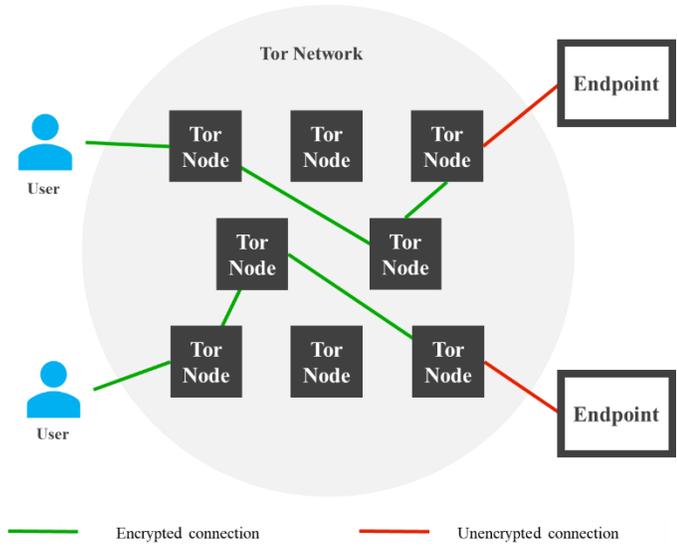


Fig. 1. Tor circuit example.

B. TLS

The encrypted communications inside the Tor network heavily relied on the TLS (Transport Layer Security). (TLS) is a cryptographic protocol designed to provide secure communication over the network [20] which could be considered as the successor of the Secure Sockets Layer (SSL) [21]. TLS is located between the Transport Layer and the Application Layer in the TCP/IP Model Layer. Other than encryption, the fundamental of TLS is the used of X.509 certificates as the verification features of with whom they are communicating, asymmetric cryptography to authenticate those entity and symmetric session key as the key to encrypt the data transfer between the entity. In addition, TLS use Record Protocol [20], which acts as a wrapper that responsible for dividing messages into several fragments. These fragments than will be paired with its corresponding Message Authentication Code (MAC). These procedures are important to ensure that the messages transferred from source to the

destination accessible by the right party and no third party could involve or do modification to these messages without the right party awareness. Other than Tor, TLS is widely use in the network environment to secure the data transfer; for example, as web browsing, email, instant messaging, and voice-over-IP (VoIP).

The knowledge on TLS protocol mechanism is very crucial in Tor traffic classification. Fig. 2 shows the Tor traffic layer. Tor will divide communication messages into several fixed size packets which called as cell [22]. Then, these cells are processed and transformed into TLS records. These records than will be fragmented into the TCP packets before it will be send to another Tor node. These traffic layer mechanism enable features extraction of the traffic at three different levels for the machine traffic classification process as each layer has certain header information that are not encrypted [33].

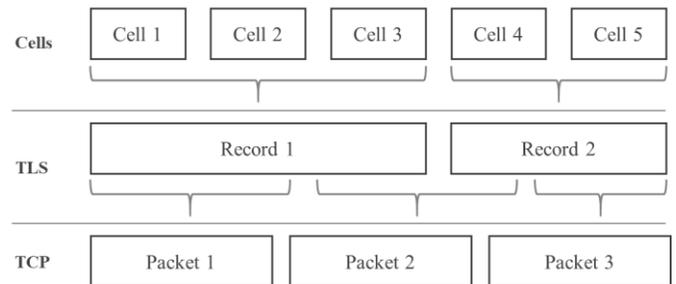


Fig. 2. Tor traffic layer [22].

III. MACHINE LEARNING

In computer science, machine learning is a technique that enable computer to learn from experience with data using statistical techniques rather than explicitly being programmed. In computer security, machine learning has been utilized in lots of area such as traffic classification, anomaly detection, spam detection, malware identification and entity classification [23].

Fig. 3 shows traffic classification taxonomy. Below is the discussion of broad categories of machine learning approaches, classification input features, classification output classes and evaluation metric of classification algorithms.

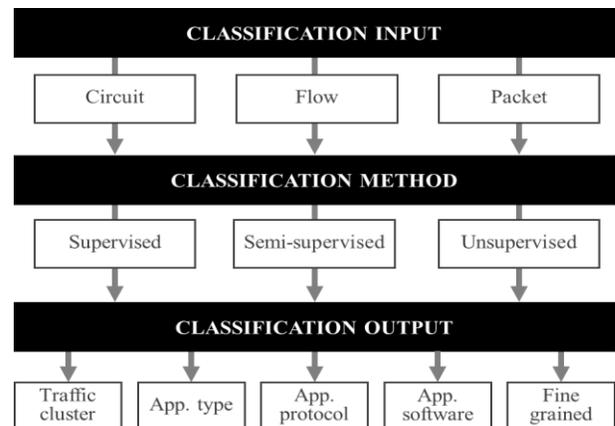


Fig. 3. Traffic classification taxonomy.

A. Machine Learning Approaches

There are three broad approaches of machine learning in traffic monitoring which are supervised, semi-supervised and unsupervised. In traffic classification, different approach produces different result, effectiveness and reliability depend on its target and the dataset that being used as the input.

Supervised learning used sets of pre-training data (data that are labelled based on certain traffic features) to train the algorithms on classification of the traffic. The example of supervised learning algorithms is K -Nearest Neighbors (k-NN) [24], Bayesian Network [25], Decision Tree [26] and Support Vector Machine (SVM) [27]. The main advantage is that it produces low false rate of classification. Yet, it might become challenging in providing complete sets of pre-training data and might require long training time.

Semi-supervised learning is similar to the supervised, but it does not utilize complete pre-training data. Instead the pre-training data is only partial labelled data. The main advantage is that there are less sets of pre-training data need to be provided. However, the accuracy might be a huge issue especially in classification various kind of encrypted application traffic.

Unsupervised learning in the other hand use mainly for data clustering without any pre-training data. The example of unsupervised learning algorithms is Fuzzy C-means [28] and K-means [29]. Despite the ease of usage as no training data is required, it commonly utilized for traffic clustering process rather than for encrypted traffic classification. One of the main usage of unsupervised traffic clustering is in anomaly detection which is work well with unlabelled traffic [29], [47], [51].

B. Classification Input

The pre-training data that will be used to train supervised or semi-supervised algorithms are fundamentally important in receiving the best output of traffic classification. Thus, selection of input data need to be done carefully and thoroughly. The Tor traffic classification input data could be segmented into three categories.

- **Circuit** – Circuit lifetime, Cell inter-arrival times, Cells per circuit life time, Uplink Cells and The Rate of the Downlink Cells to the Uplink Cells.
- **Flow** – Flow segment size, round trip time, duration.
- **Packet** – Packet length, frequency, header.

C. Classification Output

Based on the classification input, there are several types of classification output could be obtained. The best output is the fine-grained level which has the most detailed information on the classified traffic.

- **Traffic cluster (TC)** – Tor, Bulk/small transactions,
- **Application type (AT)** – Streaming, browsing, torrenting.
- **Application protocol (AP)** – HTTPS, FTP, P2P

- **Application software (AS)** – Web browser, Mail client
- **Fine-grained (FG)** – Facebook site, YouTube video, Skype call, Windows update.

D. Evaluation Metrics

As lot of researches has been carried on machine learning techniques for traffic classification, these researches need to be evaluated and compared with established findings. To evaluate the outcome of each machine learning method for traffic classification, several important metrics are used [16] which are accuracy, F-Measure, True Positive Rate (TPR), False Positive Rate (FPR), precision and recall.

IV. TRAFFIC CLASSIFICATION TECHNIQUES

This section surveys on machine learning approaches of traffic classification. The discussion will be divided into two main categories which are encrypted traffic classification that work on the Tor network and latest encrypted traffic classification on non-Tor network. Table I shows summary of cite papers and machine learning method on traffic classification.

A. Traffic Classification on Tor

Alsabah et al. [31] proposed and evaluated DiffTor that classify real-time Tor circuit using machine learning algorithms. The main intention of this study is to improve the performance of the Tor network using classification process that assigns distinct classes of services on each application traffic in the Tor circuit. Based on their observation, different applications have different time and throughput requirements. Therefore, the chosen attributes are circuit lifetime, amount of data transfer, cell inter-arrival times and number of recent cells sent are selected to classify the Tor traffic. The authors confirmed that their experiments managed to classify Tor circuit which being generated on the live Tor network with extremely high accuracy.

Similar classification technique carried out by [32] that focus on circuit and traffic flow classification. The circuit classification retrieved data at Tor's relay server and the flow classification retrieved data that are transmitting anywhere between relay server and Tor's user. This flow classification approach does not require any access on the relay server which make it much more flexible that the circuit classification approach. For circuit level classification, attributes such as cells per circuit lifetime, Uplink cells, rate of Downlink cells to Uplink cells and Exponentially Weighted Moving Average (EWMA) are chosen. For flow level classification, the authors use two flow exporting tools (Tranalyzer2 and Tcptrace) which able to generate flows and extract the attributes of the flows. The author managed to perform study on four machine learning algorithms which are Naïve Bayes, Bayesian Networks, C4.5 and Random Forest.

There is study on application classification attack on Tor network [33] which identify and classify the application types inside the Tor network traffic. The identification is based on application behaviour that characterized by traffic flow features such as burst volumes and directions. The authors define burst as successive packets in between two packets sent on the opposite direction. Based on Tor's design features

which utilize round-robin fashion for Tor scheduling, the relation between burst volumes and burst directions could be utilized as the critical features in classifying the Tor traffic. This research utilized unsupervised K-means and Multiple Sequence Alignment (MSA) to pre-treat sample data and Profile Hidden Markov Model to build model and classify the application type of the Tor traffic.

Lingyu et al. [34] proposed hierarchical classification that exploits decision tree algorithm for Tor traffic identification and Tri-Training algorithm for Tor traffic segmentation. Tri-Training algorithm is a semi-supervised machine learning algorithm which utilize co-training technique [52]. It has certain advantages such as it requires low number of training data than supervised methods, does not require cross-validation and no restriction on base classifier. Rather than focus on cell-based attributes like in [31] and [32], this study had focus on packet-based attributes which are packet length entropy, 600-byte packet frequency, zero data packet frequency (first 10) and average packet interval time. The result shows high accuracy classification which could be achieved due to the hierarchical instrument.

There is also classification study based on Anon17 dataset [30] that uses four classifier approaches (Naïve Bayes, Bayesian Network, C4,5 and Random Forest) [35]. The public dataset contains traffic from three popular anonymity services (Tor [1], I2P [3] and JonDonym [30]). The authors performed three level of classification beginning with Anon Network (Tor, I2P, JonDonym), Traffic Type (Normal, Tor Apps, I2P Apps) and Application (Tor, Streaming, Torrent, Browsing). The result of these experiment shows that all classification levels on these anonymity services could be identified and distinguished with high accuracy. There are 81 extracted features for classification including flow direction, packet length, inter-arrival time, IP header features and number of connections during traffic flow lifetime. This experiment is unique from others since it uses dataset that are publicly available. However, due to the dataset is only recently available, there lack of studies that utilize the same dataset currently.

Soleimani et al. [38] has focus on identification of Tor pluggable transports using machine learning techniques. Tor pluggable transport is a bridge from the Internet into the Tor network which considered as the technique to bypass the worldwide Tor censorship operation [45]. This experiment proceeds on three plugin techniques which are Obfs3, Obfs4, and ScrambleSuit. Using supervised learning, the identification of these plugins could be executed with only first 10-50 packets inspection in real-time. The authors utilize statistical flow features such as flow size (both direction), mean size of packet sent (both direction) and standard deviation of packet sizes (both direction).

Based on local network observer of Tor traffic dataset, [43] has analysed and found that standard HTTPS traffic (related to top monitored sites on Alexa) and Tor network has variations that could be classified using the machine learning technique. The authors generate traffic using virtual machines with two different instances. One with HTTPS traffic and the other is Tor-based traffic (both access similar website). The

authors use 40 features including total packets, total bytes, smallest packet size, largest packet size, minimum (including maximum and mean) amount of time between two packet and duration of flow. Due to the proven variation of traffic features of HTTPS and Tor network, the study outcome is a fine grained output of traffic classification (classify which traffic related to which website).

Cuzzocrea et al. [44] also presented technique that identifies Tor-related traffic that generated on a host. Hence, it could detect whether a user is using Tor application. The identification process uses supervised classification based on traffic flows features. Similar to others Tor network classification [35] and [43], the chosen attributes are 23 including flow duration, flow bytes per second, flow inter-arrival time and flow active time. This study had been carried out on six machine algorithms with the most accurate result is J48 (C4.5) approach.

B. Traffic Classification on Non-Tor Network

Fu et al. [39] has developed CUMMA, a system that use machine learning for in-App service usage classification on encrypted traffic in mobile messaging apps. This system learning model is based on temporal dependencies, user behavioural patterns and network traffic characteristics. It also works closely on time series classification and features segmentation. The traffic features extraction including packet length related features (such as descriptive statistics, variances in directions, and hopping counts) and time delay related features (such as time interval for consecutive packet). The outcome of this project shows that CUMMA enable service usage identification and application usage behaviour detection (text, picture, audio note, stream video call) based on encrypted traffic classification.

Another supervised research for encrypted traffic from [40] has proposed an attribute-aware classification that utilized second-order Markov Chain algorithm. Second-order Markov Chains is required in order to determine state transition probabilities. This study also proposes modelling process using application attribute bigram that able to increase second-order Markov Chains state diversity. The study attained better discernment accuracy and diverse application fingerprints through the leverage of the attribute bigram (Certificate and first Application Data packets). This experiment manages to classify the encrypted traffic based on the detection of website traffic-attribute which could be considered as fine grained output.

Sun et al. [41] has studied on incremental SVM (ISVM) model that focus on enabling quick and high-frequency of classifier update with reduced training cost of memory and CPU. The essential different on ISVM is that the original training data is removed and only holds the Support Vectors (SVs) produced in latest updating process which overcome the traditional SVM weaknesses. They also proposed AISVM, an ISVM with attenuation factor through the use of weight on each SVs to maximize the usage of information on SVs updating process. The outcome of this study illustrates that both proposed model shows similar classification accuracy with traditional SVM but with significant reduced updating process.

Fan et al. [36] has study on the machine learning classification that emphasizes traffic in Software Define Networking (SDN). SDN [46] is a new network paradigm that provide simplification of network management and support for exponential traffic growth on mobile cellular network. The feature selection (such as port number, number of unique data bytes, maximum segment size and initial window bytes) techniques had been use as the manipulated variable in this study as different combination of selected features produced different classification accuracy. This study employs SVM and K-means clustering for the traffic classification process. The outcome shows that K-means are effectively clustering new type of traffic; however, it has lower accuracy than supervised SVM approach.

Another SVM learning approach has been carried out by [37] that focus on real-time traffic classification. The author proposed SSP-SVM based on principal component analysis (PCA) and scaling dataset to extract and verify traffic features. It adapts the reduction on feature dimension, lower features redundancy and higher features generalization. They also utilize improved particle swarm optimization algorithm to automatically produce optimal parameter for kernel function. The outcome performance shows that two-class and multi-class classifier work effectively on traffic classification compared to the traditional SVM.

Another novel fingerprinting technique (through sampling of Application Protocol Data Units exchange patterns) for traffic classification has been proposed by [42]. The studied technique is simple to be implemented with minimal resource requirement. The principal of this technique is to provide high efficient sampling strategy that applied by single Content Addressable Memory (CAM) filtering rule based on zero-length TCP packet flows. This classification technique isn't affected with network transmission issues such as fragmentation, loses and congestion. The author also suggested that for UDP traffic, analogous fingerprinting scheme should be utilized to attain the same accuracy in TCP.

V. DISCUSSION AND COMPARISON

We have provided the summarize overview on the encrypted traffic classification using machine learning approaches. Most surveyed studies use flow and packet features as the input for the traffic classification technique. The authors of [35] manage to do classification based on both flow and packet features using public dataset [30] that are focus on the anonymity services traffic. There is also traffic classification based on circuit features [31], [32] that could be considered one of specialized machine learning process on the Tor network. This is because, the Tor circuit is only exist in the Tor network, thus it could be experimented and analysed entirely in the Tor network.

TABLE I. SUMMARY TABLE OF CITE PAPERS AND MACHINE LEARNING METHOD ON TRAFFIC CLASSIFICATION

Reference	Publication Year	Traffic Properties			Learning			Methods										Data set			Output		
		Circuit	Flow	Packet	Supervised	Semi-Supervised	Unsupervised	AdaBoost	K-Means	Naive Bayes	Bayesian Networks	Decision Trees	C4.5	Random Forest	Hidden Markov model	Tri-Training	Support Vector Machine	Second-order Markov Chain	Real-time	Public		Private	Tor compatible
[31]	2012	✓			✓				✓	✓	✓							✓		✓	✓		AT
[32]	2014	✓	✓		✓				✓	✓		✓	✓							✓	✓		AT
[43]	2015		✓		✓				✓			✓	✓			✓				✓	✓		TC,FG
[33]	2015		✓		✓		✓		✓					✓						✓	✓		AP
[39]	2016			✓	✓									✓						✓			AS,FG
[34]	2017			✓		✓					✓				✓						✓	✓	TC,AP
[35]	2017		✓	✓	✓				✓	✓		✓	✓							✓		✓	TC,AT
[36]	2017		✓		✓		✓									✓				✓			AT
[37]	2017			✓	✓											✓		✓	✓				AP
[40]	2017			✓	✓												✓				✓		FG
[44]	2017		✓		✓					✓		✓									✓	✓	TC
[38]	2018		✓	✓	✓		✓					✓	✓			✓		✓		✓	✓		AP
[41]	2018		✓		✓											✓		✓	✓				AP
[42]	2018		✓		✓							✓						✓		✓			FG

Due to the complexity of encrypted traffic, most of the surveyed research focuses on supervised machine learning which required pre-training dataset to classify the traffic. The studies by [33], [36] utilized unsupervised machine learning solely for traffic clustering rather than traffic classification. Despite study by [34] managed to classify encrypted traffic using semi-supervised learning, this learning practice could be achieved with the presence of hierarchical classification approach.

The most popular machine learning methods are the C4.5 followed by the SVM. These two methods generally produce most accurate result. However, it might not be the best algorithm on most of encrypted traffic environment. Naïve Bayes, Bayesian Networks and Random Forest also produce satisfactory accuracy when the features are adequate and reliable.

Traditional non-Tor encrypted traffic classification has the privilege of accessing important attribute such as source and destination IP address and port number to learn and classify the traffic accurately. However, in Tor network, these attributes are impractical as the source and destination info are no longer the accurate attribute as it has been hidden by the onion routing mechanism. Hence, a lot of other attributes are used for classifying the Tor encrypted network traffic.

The real-time column in Table I shows that whether the classification process could be applied on real-time data. Although accurate detection is very important, some traffic classification such as anomaly detection [47] required real-time classification process to make it useful most of the time. Hence, heavy computational and slow traffic classification is impracticable in real-time environment.

The dataset column in Table I shows the source of data whether it was publicly available or privately collected. Traffic classification on public dataset would allow researchers to closely compared their accuracy and performance result with other published work. However, due to limited availability of dataset that involve Tor network, most of the studies that focus on Tor network produced their own dataset.

Compared to the traditional machine learning research [17], current researches on encrypted traffic classification using machine learning approach produced various type of output that could be further analysed and refined. Study by [39], [40], [42], [43] managed to perform encrypted traffic classification that identify traffic with fine granularity output of information.

To this end, there is no algorithm that performs the best in all condition. Different algorithm has different capabilities and efficiency depending on its classification objective, implementation strategy and training dataset. Therefore, there are several considerations factor that need to be taken while choosing the right machine learning algorithm [9].

- **Accuracy** – Despite the most accurate algorithm is very important, the processing time might be a huge bottleneck. In some cases, approximation of traffic classification is acceptable.

- **Training time** – When data set is huge, the training time of this data might be an important consideration as it might result the training time taken to be from several minutes to few hours. Encrypted traffic classification with shorter training time is much more preferred.
- **Computational resource** – Very accurate algorithm that utilized very high computational resource might not be suitable in certain traffic classification circumstances especially in real-time classification environment. Therefore, encrypted traffic classification algorithm needs to have acceptable computational requirement for more efficient processing resources.
- **Number of features** – Commonly, the more extracted features that available, the better accuracy of encryption traffic classification. However, it might become a bottleneck on the algorithm training time process. There are also features that will reduce the accuracy of the encrypted traffic classification which need to be avoided.
- **Number of parameters** – Number that affect algorithm behavior, such as number of iterations, error tolerance and options between variants. These parameters are very important in getting the effective and accurate result since different settings could significantly impact the encrypted traffic classification outcome.

VI. CONCLUSION

In the past, traffic classification using machine learning approaches was not important. With the emerging traffic encryption and anonymity services such as Tor, machine learning technique for encrypted traffic classification should be considered as the prominent approaches on identifying this Tor traffic.

In this paper, we presented an overview of machine learning classification for Tor encrypted traffic. We begin with discussion on Tor technical background and machine learning background. Then there is discussion on surveyed papers, summarize table of the discussed papers and comparison on the machine learning approaches. To sum up, there still lots of things could be further investigated and improved in the machine learning classification process to discover the truth of privacy protection on the Tor network.

REFERENCES

- [1] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The Second-Generation Onion Router," *SSYM'04 Proc. 13th Conf. USENIX Secur. Symp.*, pp. 21–22, 2004.
- [2] Tor Project, Users – Tor Metrics, 2018. Retrieved 26 March 2018, from <https://metrics.torproject.org/userstats-relay-country.html?start=2017-01-1&end=2018-01-1&country=all&events=off>
- [3] I2P: The Invisible Internet Project. Retrieved 26 March 2018, from <https://geti2p.net/>
- [4] FreeNet. Retrieved 26 March 2018, from <https://freenetproject.org>
- [5] M. Spitters, F. Klaver, G. Koot, and M. Van Staalduinen, "Authorship Analysis on Dark Marketplace Forums," *Proc. - 2015 Eur. Intell. Secur. Informatics Conf. EISIC 2015*, pp. 1–8, 2016.
- [6] FTR Team, 2016. Cybercrime and the Deep Web. Trend Micro Security News. Retrieved 1 March 2018, from <https://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-cybercrime-and-the-deep-web.pdf>

- [7] J. Aldridge and R. Askew, "Delivery dilemmas: How drug cryptomarket users identify and seek to reduce their risk of detection by law enforcement," *Int. J. Drug Policy*, vol. 41, pp. 101–109, 2017.
- [8] J. Broséus, D. Rhumorbarbe, M. Morelato, L. Staehli, and Q. Rossy, "A geographical analysis of trafficking on a popular darknet market," *Forensic Sci. Int.*, vol. 277, pp. 88–102, 2017.
- [9] How to choose algorithms for Microsoft Azure Machine Learning, Microsoft Azure, 2018. Retrieved 30 March 2018, from <https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice>
- [10] I. Sanchez-Rola, D. Balzarotti, and I. Santos, "The Onions Have Eyes: A Comprehensive Structure and Privacy Analysis of Tor Hidden Services," *Proc. World Wide Web Conf.*, pp. 1251–1260, 2017.
- [11] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker, "Shining light in dark places: Understanding the tor network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5134 LNCS, pp. 63–76, 2008.
- [12] S. Chakravarty, G. Portokalidis, M. Polychronakis, and A. D. Keromytis, "Detecting traffic snooping in tor using decoys," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6961 LNCS, pp. 222–241, 2011.
- [13] Y. Okada, S. Ata, N. Nakamura, Y. Nakahira, and I. Oka, "Application identification from encrypted traffic based on characteristic changes by encryption," *2011 IEEE Int. Work. Tech. Comm. Commun. Qual. Reliab. CQR 2011*, 2011.
- [14] P. Mittal, A. Khurshid, J. Juen, M. Caesar, and N. Borisov, "Stealthy traffic analysis of low-latency anonymous communication using throughput fingerprinting," *Proc. 18th ACM Conf. Comput. Commun. Secur. - CCS '11*, p. 215, 2011.
- [15] P. Winter et al., "Spoiled onions: Exposing malicious Tor exit relays," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8555 LNCS, pp. 304–331, 2014.
- [16] A. Mishra, "Metrics to Evaluate your Machine Learning Algorithm", *Towards Data Science*, 2018. Retrieved 1 April 2018, from <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
- [17] P. Velan, M. Cermák, Pavel Celeda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *Int. J. Netw. Manag.*, vol. 25, pp. 355–374, 2015.
- [18] D. Goldschlag, M. Reedy, and P. Syversony, "Onion Routing for Anonymous and Private Internet Connections," *Network*, pp. 1–5, 1999.
- [19] Phobos (2010). Plaintext over Tor is still plaintext, *Tor Blog*. Retrieved 10 March 2018, from <https://blog.torproject.org/plaintext-over-tor-still-plaintext>
- [20] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2," RFC 5246 (Proposed Standard), Internet Engineering Task Force, Aug. 2008, updated by RFCs 5746, 5878, 6176.
- [21] A. Freier, P. Karlton, and P. Kocher, "The Secure Sockets Layer (SSL) Protocol Version 3.0," RFC 6101 (Historic), Internet Engineering Task Force, Aug. 2011.
- [22] A. Lazarenko and S. Avdoshin, "Anonymity of Tor: Myth and Reality," *Proc. 12th Cent. East. Eur. Softw. Eng. Conf. Russ. - CEE-SECR '16*, pp. 1–5, 2016.
- [23] R. Marty, *AI and Machine Learning in Cyber Security – Towards Data Science*, Formulated.by, 2018. Retrieved 15 March 2018, from <https://towardsdatascience.com/ai-and-machine-learning-in-cyber-security-d6fbee480af0>
- [24] S. Manocha and M. A. Girolami, "An empirical analysis of the probabilistic K-nearest neighbour classifier," *Pattern Recognit. Lett.*, vol. 28, no. 13, pp. 1818–1824, 2007.
- [25] R. E. Neapolitan, "Learning Bayesian Networks," Prentice Hall, vol. 6, no. 2, p. 674, 2003.
- [26] J. R. Quinlan (1993). C4.5: programs for machine learning. Log Altos, CA, Morgan Kaufmann
- [27] V. N. Vapnik, "An overview of statistical learning theory.," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–99, 1999.
- [28] J. Bezdek, 1981. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, Norwell, MA, USA (1981)
- [29] H. Li, "Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis," *2010 Int. Symp. Intell. Inf. Process. Trust. Comput.*, pp. 458–462, 2010.
- [30] Jondos GmbH. (2018). JonDo – the IP changer. Retrieved 13 April 2018, from <https://anonymous-proxy-servers.net/en/jondo.html>
- [31] M. Alsabah, K. Bauer, and I. Goldberg, "Enhancing Tor 's Performance using Real-time Traffic Classification Categories and Subject Descriptors," *Proc. 2012 ACM Conf. Comput. Commun. Secur.*, pp. 73–84, 2012.
- [32] K. Shahbar and A. N. Zincir-Heywood, "Benchmarking Two Techniques for Tor Classification," *Comput. Intell. Cyber Secur.*, pp. 1–8, 2014.
- [33] G. He, M. Yang, J. Luo, and X. Gu, "A novel application classification attack against Tor," *Concurr. Comput. Pract. Exp.*, vol. 27, pp. 5640–5661, 2015.
- [34] J. Lingyu, L. Yang, W. Bailing, L. Hongri, and X. Guodong, "A Hierarchical Classification Approach for Tor Anonymous Traffic," *IEEE Int. Conf. Commun. Softw. Networks*, pp. 239–243, 2017.
- [35] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescapé, "Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark," *Proc. 29th Int. Teletraffic Congr. ITC 2017*, vol. 1, pp. 81–89, 2017.
- [36] Z. Fan and R. Liu, "Investigation of machine learning based network traffic classification," *2017 Int. Symp. Wirel. Commun. Syst.*, pp. 1–6, 2017.
- [37] J. Cao, Z. Fang, G. Qu, H. Sun, and D. Zhang, "An accurate traffic classification model based on support vector machines," *Int. J. Netw. Manag.*, vol. 27, no. 1, pp. 1–15, 2017.
- [38] M. H. M. Soleimani, M. Mansoorizadeh, and M. Nassiri, "Real-time identification of three Tor pluggable transports using machine learning techniques," *J. Supercomput.*, pp. 1–18, 2018.
- [39] Y. Fu, H. Xiong, X. Lu, J. Yang, and C. Chen, "Service Usage Classification with Encrypted Internet Traffic in Mobile Messaging Apps," *IEEE Trans. Mob. Comput.*, vol. 15, no. 11, pp. 2851–2864, 2016.
- [40] M. Shen, M. Wei, L. Zhu, and M. Wang, "Classification of Encrypted Traffic with Second-Order Markov Chains and Application Attribute Bigrams," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 8, pp. 1830–1843, 2017.
- [41] G. Sun, T. Chen, Y. Su, and C. Li, "Internet Traffic Classification Based on Incremental Support Vector Machines," *Mob. Networks Appl.*, pp. 1–8, 2018.
- [42] J. Kampeas, A. Cohen, and O. Gurewitz, "Traffic Classification Based on Zero-Length Packets," *IEEE Trans. Netw. Serv. Manag.*, vol. 4537, no. c, pp. 1–14, 2018.
- [43] A. Almubayed, A. Hadi, and J. Atoum, "A Model for Detecting Tor Encrypted Traffic using Supervised Machine Learning," *Int. J. Comput. Netw. Inf. Secur.*, vol. 7, no. 7, pp. 10–23, 2015.
- [44] A. Cuzzocrea, F. Martinelli, F. Mercedo, and G. Vercelli, "Tor Traffic Analysis and Detection via Machine Learning Techniques," *IEEE Int. Conf. Big Data*, pp. 4392–4398, 2017.
- [45] Tor Project, *Tor: Pluggable Transports*, 2018. Retrieved 15 April 2018, from <https://www.torproject.org/docs/pluggable-transports.html.en>
- [46] ONF, *Software-Defined Networking (SDN) Definition - Open Networking Foundation*, 2018. Retrieved 15 April 2018, from <https://www.opennetworking.org/sdn-definition/>
- [47] S. Omar, A. Ngadi, and H. H. Jebur, "Machine Learning Techniques for Anomaly Detection: An Overview," *Int. J. Comput. Appl.*, vol. 79, no. 2, pp. 975–8887, 2013.
- [48] Z. Chen, L. Ruan, J. Cao, Y. Yu, and X. Jiang, "TIFAflow: Enhancing traffic archiving system with flow granularity for forensic analysis in network security," *Tsinghua Sci. Technol.*, vol. 18, no. 4, pp. 406–417, 2013.
- [49] Y. Wang, Y. Xiang, J. Zhang, W. Zhou, G. Wei, and L. T. Yang, "Internet traffic classification using constrained clustering," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2932–2943, 2014.

- [50] S. S. L. Pereira, J. L. De Castro E Silva, and J. E. B. Maia, "NTCS: A real time flow-based network traffic classification system," Proc. 10th Int. Conf. Netw. Serv. Manag. CNSM 2014, pp. 368–371, 2015.
- [51] N. Goernitz, M. M. Kloft, K. Rieck, and U. Brefeld, "Toward Supervised Anomaly Detection," J. Artif. Intell. Res., vol. 46, pp. 235–262, 2014.
- [52] Z.-H. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers," IEEE Trans. Knowl. Data Eng., vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

The Implementation of Computer based Test on BYOD and Cloud Computing Environment

Ridi Ferdiana

Electrical and Information Engineering Department
Universitas Gadjah Mada
Sleman, Indonesia

Obert Hoseanto

Sampoerna University
Jakarta, Indonesia

Abstract—Computer-based test promises several benefits such as automatic grading, assessment features, and paper efficiency. However, besides the benefits, the organization should prepare the enough infrastructure, network connectivity, and user education. The problem upsurges when a hundred numbers of users join the computer-based test. This article proposes Bring-Your-Own-Devices (BYOD), and the cloud computing approach to facilitate a hundred numbers of exam participant. Through the experiment method on 393 students, the article determines five central practices that can be used by the organization who want to implement the massive scale computer-based test.

Keywords—Computer-based test; cloud computing; Bring-Your-Own-Devices (BYOD)

I. INTRODUCTION

In a developing country like Indonesia, adopting a computer-based test (CBT) in high school is challenging. The CBT lab preparation, lab policy, student learning challenges, and staffing preparation [1]. The preparation is not free either, for example; the infrastructure should be prepared and invested such as the internet, computer, and network. The problem is when the infrastructure cannot be achieved.

One way to achieve the infrastructure is by adopting Bring-Your-Own-Device (BYOD). It can be used to support learning activities such as CBT in high school or higher education [2]. In the implementation of the CBT, many of students prefer to use the mobile phone to access the CBT [3]. To prepare the student to get accustomed to the CBT system, several types of research do:

- The development of mobile learning system to make student accustomed to the written exam CBT [4].
- The automated of course management and evaluation system using mobile web [5].
- The development of an android mobile application for e-exam [6], [7].
- The knowledge level assessment in e-learning system that uses machine learning and user activity analysis [8].

Based on the researches, it is shown that the solution to implement CBT with BYOD is by creating a custom solution to facilitate the CBT. However, in the limited resources such as a high school that doesn't have a programmer or IT

administrator. The solution is somewhat tricky. Hardware and software modularity, cost-effectiveness, network, and training are the main factors to implement cloud computing [9]. Therefore, the solution that possible is by utilizing cloud computing. In the recent researches, it is shown that cloud computing has the potential to provide simulator facility or ICT enhanced classroom [10], [11]. Some researchers also suggest using cloud computing through a teacher training program [12]. Although the implementation of cloud computing on education seem promising and useful [13], some questions need to be answered:

- Is the combination between BYOD and Cloud Computing can be implemented on a real case?
- What are the critical factors that need to be considered to implement BYOD and cloud computing?

The questions are referring to the problem to run the CBT with efficient cost and less technical effort. In order to answer the question, this article conducts an experiment to implement BYOD and cloud computing in the CBT. The detail of the experiment is described in Section II.

II. RESEARCH DESIGN

The research design follows qualitative research design on information system [14]. It uses case study design by doing participant observation in natural settings to explain the implementation of BYOD on CBT (Table I). The qualitative method follows structured interview that is documented and administered through online questionnaires.

Several considerations in this research are:

- 1) The research uses student on eight grades rather than seven grades or nine grades. The middle grade is chosen to balance the digital literature knowledge.
- 2) The CBT software uses Microsoft Forms. It is well integrated on Microsoft Office 365 that already used widely in the school. It is to make sure that the teacher can author their own exam and troubleshoot as a proctor on the exam.
- 3) The bandwidth that is allocated uses a wireless network provider. The selection of the provider based on the fastest wireless coverage of that era.
- 4) The school is located on the Jakarta. The author selects the school based on the school agreed to adopt BYOD on CBY. To mimic the majority school in Indonesia, the author

chooses a school that doesn't have dedicated hardware for the CBT process.

TABLE I. RESEARCH DESIGN

Research design attribute	Values
Design Groups	Case study
Research Method	Qualitative
Technique	Structured
Instrument	Questionnaires
Research Object	Exam participants on Junior High School (Grade 8)
Population	396 students on grade 8
Samples	393 students
CBT Software	Microsoft Forms (parts of Office 365 SaaS)
Software Cost	Free (for the academic institution)
Experiment Length	Four days (14 March – 18 March 2018)
Classroom	22 Classroom simultaneously
Bandwidth	Shared Bandwidth 35 Mbps downstream / 4 Mbps upstream
Network Distribution	Through 22 Access point (one classroom - one access point)
CBT experiment	8 experiments (8 subjects)

The research was done on four steps, namely, infrastructure preparation, exam preparation, exam execution and monitoring, exam evaluation.

III. IMPLEMENTATION PROSES AND RESULT

The result will be described in four steps as mentioned in the research design.

A. Infrastructure Preparation

Infrastructure preparation prepares the infrastructures needed to run the CBT on BYOD system. The infrastructure preparation is done by doing assessments through simulating the exam process in the cloud environment. This step is done by following three main steps.

1) Creating questions on exam software such as Microsoft Forms. On this step, the teacher creates four question configurations as shown in Table II.

2) Measuring the demand estimation for the users. The demand estimation is measured by using network traffics feature in the browser developer toolbar. The demand estimation is shown in Table II.

TABLE II. EXAM CONFIGURATION

Question Configurations	Content Composition	Size	Demand Estimation for 432 users
40 questions	50% image 50% text	2.1 MB	907.20 MB
20 questions	50% image 50% text	1.2 MB	518.4 MB
20 questions	30% Image 70% text	533 KB	230.3 MB
20 questions	100% text	58.2 KB	25.2 MB

3) Based on Table II, the required bandwidth is 8 Mbps. Assuming that the traffic workload is 25% for the CBT. It will need at least 32 Mbps.

4) The economic router will only support 16 clients. 386 students need 18 routers (386/16). However, because of the physical classroom, the research uses 22 routers to facilitate 22 classrooms.

B. Exam Preparation

The exam preparation process is done by creating an exam question on several subjects. There are eight subjects for four days. The subjects are Mathematics, Science, Social, Religion, Indonesian Language, ICT, Arts, and Society. Each subject has 40 questions. The questions are created by the teachers and uploaded to Forms.

After the questions are uploaded, the questions are verified by the exam preparation team. The preparation will evaluate the exam based on several rules such as:

- The question and answer should be in a text format, except for the figure or the tables.
- The question should be in multiple choice format
- The figure and the table should be readable on mobile devices

The validated question is uploaded and shared as a specific link. The link is converted into QR codes. The link can be accessed at specific time and date. The link is evaluated through a load test. The load test is done by following three steps which are creating a test plan, recording the usage scenario through Visual Studio UI coded test, preparing the cloud test environment to test the UI coded test based on the test plan. The test plan consists of several rules:

- There are two types of questions set. The lightweight exam sets and heavyweight exam sets. The lightweight exam sets consist of 100% of the exam is test while the heavyweight exam sets have a picture on each question. The number of the question is 40 questions.
- The length of the test is 30 minutes, 60 minutes, and 90 minutes. This is based on how the student will finish the 40 questions.
- The load test will be done for 400 users with 100 iterations using Microsoft Azure.

C. Exam Execution and Monitoring

Based on the load test, there are several ways to implement the CBT process on the cloud which are:

- The exam process is executed by giving each classroom a delay between 5-10 minutes. This is to make sure that the workload of the network still sufficient to download the exams.
- The devices should free from an update such as Windows update, Android update, or apps update.
- The exam should begin earlier to give some time for the student to turn off an unused app that consumes

bandwidth, turn on the wireless module to get connected and download the exam. It will take five to ten minutes. Therefore, the student should prepare the devices 30 – 60 minutes before the exam.

- The student starts the exam by using the QR codes to access the exam portal.
- The student will authenticate through Office 365 account that distributed a week before the exam. The student should test and use the Office 365 especially Microsoft Forms before they use it for the exam.
- The need of Proctor in each classroom and one technical support for the exam solution.
- The indirect assessment of BYOD implementation is used through an exit survey. The student will answer the survey question in Table III. The survey is done through an online mechanism.

D. Exam Evaluation

After the implementation, it is found several challenges and phenomena that can be discussed as follows:

- The upstream issues. The student should wait between 1 – 5 minutes to submit the answer to the exam. On the tested environment, the downstream is on 30 Mbps while the upstream is on 10 Mbps.
- International bandwidth usage. The Microsoft Forms uses Office 365 that hosted on the Cloud. Therefore, the system uses international bandwidth rather than local bandwidth.
- Difference use experience on the device. The students use a variety of devices started from a smartphone with 4' screen to 15' laptop screen. The variety makes the student should calibrate the zoom factor and answering process.

The exit survey is done when a student finished eight subjects on four days. In the last exam, the student should answer the survey questions. The survey is distributed on 396 students, with 393 valid samples. The average time to complete for the survey is 3.46 minutes. Some finding of the exit survey that related with the cloud computing adoption for BYOD exam solutions are:

- 1) The students satisfied with the BYOD exam solution with cloud computing. (4.26 of 5.00)
- 2) The students feel excited about the BYOD exam solution. (4.28 of 5.00)
- 3) The students have a good exam experience (4.27 of 5.00).
- 4) 82.2% of students agree that the next exam should use CBT exam.
- 5) 62.2% of students agree that any exam can be converted into online exam experience.
- 6) 61.8% of students agree that the capability of digital literacy helps them to finish the exam without a problem.
- 7) 60.5% of students agree that the CBT exam has no problem at all. While other students don't like the concept

because slow internet connection (25.3%) unsupported device (3.4%), and user experiences issues on various screen (10.9%).

8) 61.2% of the students feel that the CBT exam has no difficulty at all. While the 19.6% said an easy task, 1.6% said really easy, and 17.3% students said that the exam software is difficult to operate.

9) 8.3% of students said that the internet connection is too slow. While 43.2% feel satisfied and 48.6 feel sufficient.

10) 92.5% of students use their own devices while 7.5% uses the device from the school.

11) 34.1% of students access the internet every day. While the 38.8% access on a weekly basis, 13.2% monthly basis, and 14% for bi-weekly basis. This background will help the digital literacy level for the students.

12) 94.6% of students use smartphone, 3.9% uses notebook, and 1.5% uses tablet. This phenomenon shows that the most used BYOD device is a smartphone.

13) 81.9% of students said that the CBT exam is not hard to understand. 18.1% students feel that the CBT exam is really hard to uses.

14) 65.2% of students said that the success of BYOD is mostly because of internet connection, 14.47% said the sufficient hardware, and 11.63% is about the socialization of the software. The rest students feel that the combination of the three is the success factor of BYOD exam.

Based on the exit survey, it is shown that the BYOD implementation for the computer-based test with the cloud computing can be adopted by considering several critical factors such as internet connection, the client device/hardware, the exam execution process, student digital literacy, and ready to use software to do CBT like Microsoft Forms.

IV. CONCLUSIONS

This paper discusses the field experiment of the BYOD implementation in the cloud environment for a computer-based test. As a result. The computer-based test can be used by using BYOD and cloud with five main practices that can be a consideration:

1) Capacity planning should be done to understand the usage pattern of the computer-based test. The usage pattern can be simulated through a cloud load test, coded-UI test, and network test.

2) The three key factors that should be prepared before the exam is the readiness of the internet connection, the supported hardware and the socialization of the software to the students.

3) The exam processes should be prepared such as the device validation, the use of QR codes, and the delay time between classrooms to download the exam.

4) The BYOD implementation should consider user experience for the Smartphone, Notebook, and the Tablet.

5) The Cloud-based CBT should consider the bandwidth consumption between international or local bandwidth, upstream and downstream, and a total of wireless access coverage that should be prepared for simultaneous access.

ACKNOWLEDGMENT

We would like to thank Microsoft Innovation Center UGM for cloud computing infrastructure. The research is supported by the Department Electrical and Information Engineering Universitas Gadjah Mada.

REFERENCES

- [1] Y. Tunc and M. Armstead, "Computer-based testing," in Proceedings of the 29th annual ACM SIGUCCS conference on User services - SIGUCCS '01, 2001, p. 201.
- [2] H. Berger and J. Symonds, "Adoption of Bring Your Own Device in HE & FE Institutions," in Proceedings of the The 11th International Knowledge Management in Organizations Conference on The changing face of Knowledge Management Impacting Society - KMO '16, 2016, pp. 1–6.
- [3] Á. Matthíasdóttir and H. Arnalds, "e-assessment," in Proceedings of the 17th International Conference on Computer Systems and Technologies 2016 - CompSysTech '16, 2016, pp. 369–374.
- [4] L. Horvat, J. Balen, and G. Martinovic, "Proposal of mLearning system for written exams," in Proceedings ELMAR-2012, 2012, pp. 345–348.
- [5] M. S. A. El-Seoud, A. Karkar, I. A. T. F. Taj-Eddin, H. F. El-Sofany, A. Dandashi, and J. M. Al-Ja'am, "Semantic-Web automated course management and evaluation system using mobile applications," in 2015 International Conference on Interactive Collaborative Learning (ICL), 2015, pp. 271–282.
- [6] R. Nagal, P. Nemkul, D. Kumar, N. Kumar, and A. Joseph, "Android based Secure Exam Management System to Prevent Impersonation," Int. J. Latest Technol. Eng. Manag. Appl. Sci., vol. VI, pp. 46–49, 2017.
- [7] L. Dan Cheng and X. C. Wang, "Mobile application tools for learning and quiz based on Android," in 2013 IEEE 63rd Annual Conference International Council for Education Media (ICEM), 2013, pp. 1–1.
- [8] N. Ghatasheh, "Knowledge Level Assessment in e-Learning Systems Using Machine Learning and User Activity Analysis," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 4, pp. 107–113, 2015.
- [9] A. M. Kadhum and M. K. Hasan, "Assessing the Determinants of Cloud Computing Services for Utilizing Health Information Systems: A Case Study," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 2, p. 503, Apr. 2017.
- [10] V. Tam, A. Yi, and E. Y. Lam, "Building an Interactive Simulator on a Cloud Computing Platform to Enhance Students' Understanding of Computer Systems," in 2013 IEEE 13th International Conference on Advanced Learning Technologies, 2013, pp. 154–155.
- [11] T. Yamanoue, K. Oda, S. Tetaka, and K. Shimozono, "Portable cloud computing system - A system which makes everywhere an ICT enhanced classroom," Proc. ACM SIGUCCS User Serv. Conf., pp. 85–88, 2014.
- [12] Y. Kim, "The Framework of Cloud e-Learning System for Strengthening ICT Competence of Teachers in Nicaragua," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 1, p. 62, Feb. 2018.
- [13] Q. A. Alajmi, A. Kamaludin, R. A. Arshah, and M. A. Al-Sharafi, "The Effectiveness of Cloud-Based E-Learning towards Quality of Academic Services: An Omanis' Expert View," IJACSA Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 4, 2018.
- [14] M. A. Taylor, Information System Research. Momentum Press, 2017.

Method for Designing Scalable Microservice-based Application Systematically: A Case Study

Ahmad Tarmizi Abdul Ghani, Mohamad Shanudin Zakaria
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Malaysia

Abstract—Microservice is a new transformation of Service-Oriented Architecture (SOA) which is gaining momentum in both academic and industry. The success of microservice began when giant companies like Netflix used them as a service architecture for the purpose of serving customers. Monolithic architecture used by Netflix previously was no longer able to cope with business development and it is difficult to scale to meet user demands. Although Netflix has been successful with microservice architecture, there is no systematic method introduced to produce microservice. Academic studies related to microservice are still in the early stages and have not yet reached maturity. Microservice is seen to require a method that helps organizations to systematically design microservices and replicate the success achieved by Netflix. In forming a method for this systematic microservice then the methods for building an existing microservice are studied. Based on the Design Science Research method, two research artefacts have been produced. The first artefact is a systematic design of microservice that has four main steps. The second artefact is the instantiation by applying the proposed microservice design method to the case studies, namely, MyFlix. Next the evaluation is made on the new method produced by obtaining expert opinions through the process of demonstration and interviewing. The expert assessment results found that the proposed method was able to produce a systematic microservice design based on the six proposed principles and the four main steps. This method can also produce a complete feature microservice such as cohesive, loose coupling, distributed and decentralized that will contribute to the production of scalable system.

Keywords—Microservice; systematic method; scalable microservice design

I. INTRODUCTION

The company or organization today has an information system that has been operating for a long period of time. This system is constantly evolving and in line with changes in business processes within the company/organization. Upon reaching one level, the system can no longer grow due to physical constraints. Such a system is known as a monolithic system. The monolithic system is an undeveloped and centralized architecture. Such a system is a vendor locked-in system and requires a high cost for system upgrades to meet the increasing system requirements [2]. Although SOA has been introduced to address the problem of isolated system integration, SOA is dominated by large companies such as Oracle, IBM, and Microsoft that produce SOA in the box [2]. SOA in such a box is known as ESB and it is vendor locked

and charges high cost to scale. ESB is a monolithic SOA architecture.

Companies that do not use SOA ESB however still use client-server-based (n-tier) systems. Customer-based systems are servers that are centrally and physically constrained. This is what happened to companies like Netflix, Uber and Zalando originally built a client-based system that eventually became a monolithic system. Netflix, Uber and Zalando are examples of ever-expanding companies. The company is a company that provides services to customers online. The increase in the number of users will increase the amount of access to the company's system. Therefore, the system needs to be scaled to meet increased use. If this is not done then the company cannot provide the best service to the customer (for example, the service becomes slow or inaccessible). The adverse effects on user experience will cause the user to quit subscribing to the services provided by these companies. To address the monolithic scaling constraints, organizations or companies need to exit ESB architecture and client-server architecture and find solutions to the problem. The first Netflix has taken steps to scale their systems beyond the client-server architecture. Microservice architecture has been used to develop Netflix systems that can be scaled without physical constraints. As a result of Netflix's success, microservice have begun to get attention from companies and organizations that have similar problems like Netflix. Netflix's success is welcomed by the creation of many tools to build microservices run by Spring Boot, Java EE, Prometheus, Hystrix as well as cloud technology providers such as Google Cloud, Amazon Lambda and Microsoft Azure. The framework for developing this microservice is however a tool that simplifies the transformation to a microservice-based system that is specific to a particular language for example Spring Boot is only specific to Java users. The framework is also a microservice architecture implementation that emphasizes only the aspects of microservice infrastructure such as load balancing, gateway APIs, microservice registries, microservice monitoring over functional aspects of microservice-based applications.

II. RESEARCH BACKGROUND

There are several methods that have been developed for the purpose of producing microservice-based applications. Among them are Domain Based Design Methods [8], [13], [18], Choreography Methods [19], User Interface Based Methods [17], Organization Structure Based Methods [6], [10], Service Cutting Techniques [11], Microservice Dependency

Engineering [16] and Microservice Extraction Techniques [12]. The methods studied have not been emphasized by the end result of a scaled-down microservice application design that results in an cohesive, loosely-coupled, distributed and decentralized [9]. Methods of transforming monolithic applications into microservice-based applications are also found to be part of the information technology division, while the best way to design microservice-based applications should take into account organizational and business aspects [9]. The methods used are still not systematically processed. Systematic method means a method with a system. Systematic methodology means it has a step-by-step guide that must be made by the developer to produce the desired design [14]. Each step has a clear and detailed explanation of the required input, the process to be done and the resulting output. Systematic methods are important because they can guarantee the development of scalable microservice-based applications instead of producing monolithic applications that are not scalable.

III. RESEARCH METHODOLOGY

The research method used in this study is the Design Science Research (DSR). This method has been used as this study is in the form of design science. Design science uses scientific methods to analyse the structure of the technical systems and its relationships with the environment. It also aims to publish rules in developing systems using the system elements and their relationship [14]. The output of the DSR study is construct-shaped artefacts, models, methods and instantiations. The main output generated from this method is a method of designing a systematic microservice-based application. The following are the phases in the DSR study method used for this study [15]:

A. Problem Identification and Motivations

The research problem has been identified by using Systematic Literature Review (SLR). All the key papers regarding methods and techniques for transforming monolithic system into microservice has been reviewed.

B. Defining the Objective to a Solution

Once research problem has been identified, the objective of the research also need to be defined. In order to solve problems in existing microservice design method, the objective has been set to design a method that can be used to design scalable microservice-based application systematically.

C. Design and Development

Based on the objective to develop a method that can be used to design scalable microservice-based application systematically, the design principles need to be identified first. Then, steps of the method are developed based on the principles.

D. Demonstration

Method developed then has been applied to a case study and then being demonstrated to the experts. The experts are the software developer with more than 10 years of experience. The method used for the demonstration was walkthrough method.

E. Evaluation

The method then evaluated by the experts. There are 5 experts have been interviewed. The experts were first demonstrated with the walkthrough of the application of the proposed method to a case study. The experts then are being interviewed with questions regarding the effectiveness of the method in designing scalable microservice-based application.

F. Communication

Results from the research is then reported in publication to disseminate the knowledge about the proposed method.

IV. METHOD FOR DESIGNING SCALABLE MICROSERVICE-BASED APPLICATION SYSTEMATICALLY

The proposed method for designing a scalable microservice-based application systematically consists of four major steps (Fig. 1):

A. S1: Model the Organization Unfolding Structure

Step S1 is modelling the organization's unfolding structure. The organization's unfolding structure should be modelled before any other steps can be implemented. The resulting structure represents the actual organizational structure of a company or system. The organizational structure of a company described here does not mean a hierarchical organizational structure but a viable organizational structure [1]. An organization is also a system where the system definitions here are interconnected and working together as an entity [7]. So, in order to control a system then the model should be produced because "every good controller for a system is a model for the system" [5].

B. S2: Model the Business Process

The second step is to model the business process after the model of the organization's unfolding structure/system has been created. The unfolding structure model allows organization/system transformation activities to be identified. This transformation activity means any activity to convert input to output. This activity is also known as the main activity of the organization [7]. Key identified activities will then be mapped to the relevant business processes. Thus, an organizational business process should also be modelled and mapped to the main activities. The difference between the business process and the main activity is the responsible business process of the value chain whereas the main activity is in regards to regulation [7]. The business process can be modelled in many ways. One of the methods chosen in this proposed method is to model with the promise theory. The promise theory is used as it is a new approach in building a distributed and autonomous system [3]. Promise theory emphasizes equilibrium in building a distributed system. It can also prevent the occurrence of situations where dynamics override semantics where conditions can be controlled to a dynamic system and maintain system stability [4]. Since microservices are distributed systems then modelling business processes with promise theories can generate autonomous and centralized microservice and it separates between intent and implementation.

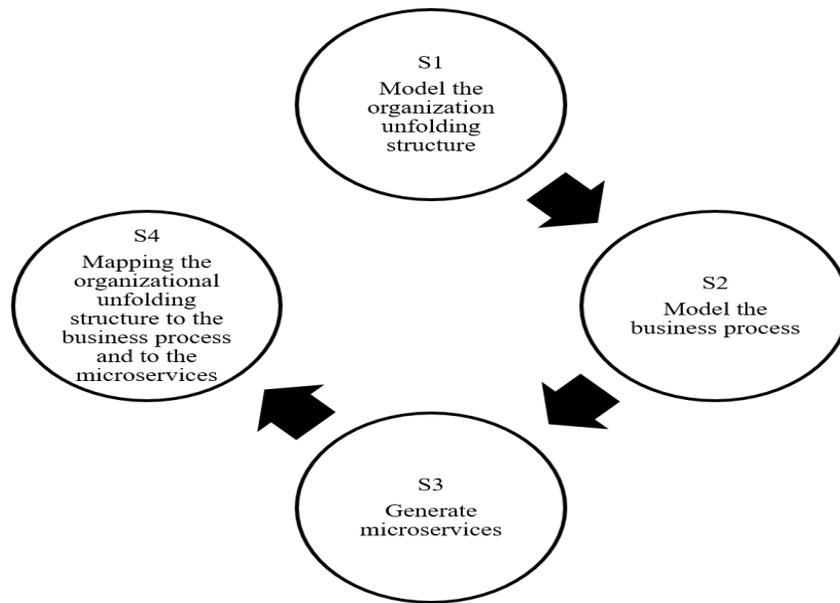


Fig. 1. Steps in the proposed method.

C. S3: Generate Microservices

Based on the business process model using the promise theory that results in the S2 step then the microservice candidate can already be generated. Generating microservice involves earning the name of the microservice candidate as well as the Application Programming Interface (API). APIs are channels for microservice to service users. It is similar to the function in programming that is the function name, parameter/input to the function and output to be generated by the function. The resulting microservices and APIs are not specific to any programming language and they aim to make the resulting microservices implementable in any programming language.

D. S4: Map the Organization's Unfolding Structure to the Business Process to Microservices

The last step is to update the unfolding structure. Artefacts generated from the S2 steps (business process) so that S3 (microservice) should be placed together with the technology branch in the unfolding structure. The unfolding structure needs to be updated to ensure that the unfolding structure is up to date as it will be a reference to any party wishing to subscribe to the existing microservice.

The principles behind the design method are as follows:

1) Principle 1: Systematic

In order to achieve the desired results then there should be a systematic method of systematic methodology, demonstrating step by step in a clear and detailed manner resulting in the desired results [14]. This proposed method emphasizes the creation of systematic design of microservice-based applications so that the desired design results are consistent with the characteristics of a microservice-based application primarily from a distributed and decentralized system aspect.

2) Principle 2: Reducing complexity

System construction is a complex process. The proposed methodology is not to increase the complexity of system

construction but to reduce complexity to a degree that can be controlled by system developers [7]. Similarly, the resulting microservice-based application will be a more organized and structured system.

3) Principle 3: Balance

The system based on the microservice generated through this method will be more stable in terms of semantic and dynamic [4]. The paradigm is taken from the promise theory of the importance of the system interacting voluntarily as opposed to coercion. Microservice-based applications need to be modelled with promise theories in order for the system to achieve a balance between the services that interact with each other. It also makes the microservice more reliable.

4) Principle 4: Documentation

The measures taken in this method will produce artefacts that are also documentation of developed microservice-based applications [4]. Documentation is especially important as it will be a reference or a trail to developers who develop the system by himself or a new developer trying to understand the system.

5) Principle 5: The alignment of organizations, businesses and information technology

Microservice-based applications aim to enable the system to evolve in parallel with changes in businesses and organizations [7]. Businesses are changing so fast that they need to meet the needs of new customers. Changes to organizations and businesses mean there will be changes to the business processes that need to be supported by information technology. This proposed method allows microservice to be produced more quickly, less complex, systematic and also stable to meet evolving business needs.

6) Principle 6: Scaled microservice design

The main feature in microservice production is to produce a scalable, cohesive, loosely-coupled, distributed and decentralized service [9].

- S1. Model the organization's unfolding structure from complexity drivers
 - S1.1 Model organizational technology structure
 - S1.2 Model the organizational geographical structure
 - S1.3 Model the organizational client/provider structure
 - S1.4 Model the organizational structure of time
 - S1.5 Combine the models from step S1.1 - S1.4 to become the organization's unfolding structure
- S2. Model the business process
 - S2.1 For each branch from the root of the organization's unfolding structure resulting from S1.5
 - S2.1.1 Start from the bottom level technology branch n to level 0
 - S2.1.1.1 For each level, model the business process for each branch of technology at that level
 - S2.1.1.1.1 If the level is the leaf level, model the business process by taking into account all types of complexity drivers that will use it at the higher level
 - S2.1.1.1.2 If it is not a leaf level, model the business process by taking into account all types of complexity drivers that will use it at the higher level and all types of complexity drivers used at the lower level
 - S2.2 Convert the resulting business process into the form of promise statement
 - S2.3 Model the promise diagram based on promise statement from step S2.2
- S3. Generate microservices
 - S3.1 List the microservice candidates from step S2.3
 - S3.1.1 Determine the API for each microservice candidate based on the promise model from step S2.2
 - S3.2 Model a microservice dependency diagram
- S4. Map the organization's unfolding structure to the business process to microservices

Fig. 2. Algorithm of the Proposed Method.

The algorithm for designing the scalable microservice-based application (Fig. 2).

V. CASE STUDY

Chief Technology Officer at F.com (real company name has been hidden) stated the company's desire to develop an application called MyFlix to download videos and watch videos online to paid subscribers. MyFlix applications must be developed using microservice architecture as a company that provides the similar services Netflix has adopted microstructure architecture to develop their system. The company chooses microservice because the architecture is seen capable of being properly scaled and suited to use cloud technology for scaling and also load balances. This service is expected to be supplied to customers around the Southeast Asia region from Malaysia and Singapore. This system should be designed so that it can accommodate high demand at one time without affecting service delivery to the user.

The customer will use the service by becoming a member and making monthly payments through the bank account. The business model used is to serve customers and receive monthly payments is a familiar model for online services today. The process to enable users to register and get the service should be smooth so that users can access the system for immediate service.

The service provided by this company is a search and download service online using cloud technology, stores the video in cloud technology and users can watch the video directly from the internet without having to download video files into their computer or mobile device.

The services provided are intended to enable users to quickly download video files from video providers because cloud technology has high download capability compared to downloading using custom broadband. The user can also download as many videos at the same time and the system will behave on behalf of the user without the user waiting for each download ready. The system also needs to provide the facility to delete video files that are no longer needed. Another feature is the convenience of searching video files on the Internet and continuously downloading them into the system. The system also needs to convert the video format into a format that can be displayed on a web browser or mobile device.

The load on the system should be propagated and better balanced to prevent any failure on the system. The system will later be seen primarily on search, downloading and viewing videos. Therefore, these critical parts need to have a good load balancing and scaling strategy so that the service can continue to be supplied to the user at a user acceptable rate.

Developing a new system is especially risky, especially in this case providing online services that involve a high number of users, a wide range of tools used, users from multiple locations as well as peak times where users access multiple systems. Each one can affect the supplied service. If the system is centrally developed by putting all the services in a single server then there is definitely a service that gets high demand will affect other services. Important service segregation so service that has high demand will not affect other service. Building a microservice-based application requires a different approach than a centralized and non-distributed traditional approach. It requires methods and approaches in which the system can be designed to produce distributed and decentralized services and is able to provide developers with an understanding of the behaviour of the built-in microservice system. It requires a balance between semantics and dynamics.

A. S1: Model the Organization's Unfolding Structure from Complexity Drivers

To simplify the S1 step implemented then the software tool has been used. The software used is Treeline which is the software to generate the main diagram using XML. By using this software all sub steps in S1 can be done directly by using the software. To make it easier for users to replicate the tree structures generated using the software then it also includes an XML schema for each tree-producing step. Next, users can modify tree structures such as adding, updating and deleting to produce an ideal tree structure.

B. S1.1: Modeling Organizational Technology Structure

The model of the MyFlix technology model that has been produced (Fig. 3). The model illustrates MyFlix's organizational structure/system structure comprising three main activities (technology), Membership Services, Browsing Services and the Download Service. Every major activity has sub activities as in Membership Services containing Registration and Payments, Listed Videos Available Video List, Play Video, Download Video Files and Delete Video and Download Services comprises Search technology, List Search Results, Create Playlist, Download and Change Format .

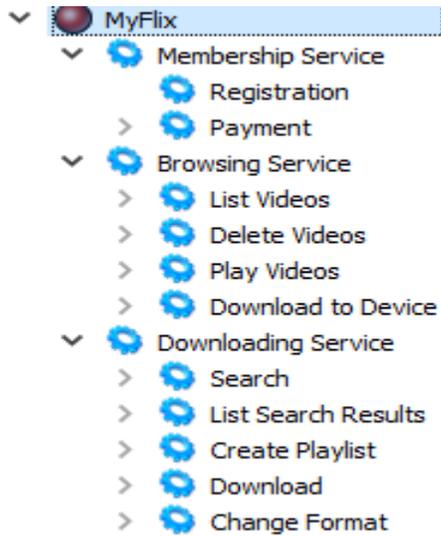


Fig. 3. MyFlix technological model.

C. S1.2: Modeling Organizational Geography Structure

The model of the MyFlix geographical that has been produced (Fig. 4). The model describes the geographic structure of the MyFlix organization/system comprising a geographic drive South East Asia. South East Asia consists of two sub geographies namely Malaysia and Singapore. The geographic drive explains the context where MyFlix's service/technology will operate.

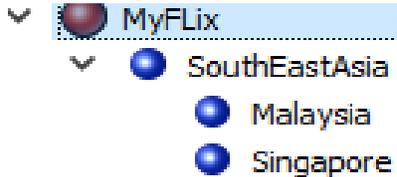


Fig. 4. MyFlix Geographical Model.

D. S1.3: Modeling of Client and Organization Provider Structure

The model of the client tree and the supplier of MyFlix (Fig. 5). The model illustrates the structure of the MyFlix organization/system client consisting of two types of client i.e. Free and Paid. MyFlix's organizational structure/system is comprised of four types of suppliers, Free Video, Paid Video, Shared Video and Subscribed Video.

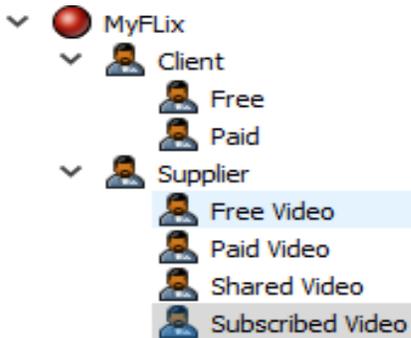


Fig. 5. MyFlix client and supplier model.

E. S1.4: Modeling Organizational Time Structure

The MyFlix tree time model that has been produced (Fig. 6). The model describes the structure of the organization's MyFlix system comprising years and followed by months. This year is the year of MyFlix operation beginning in 2018. Sub year time consists of 12 months from January to December.

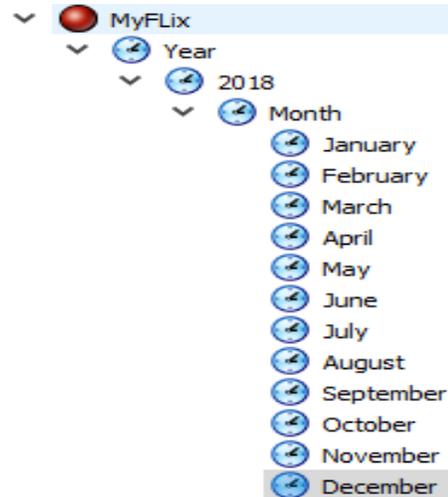


Fig. 6. MyFlix time model.

F. S1.5: Integrate Models from Step S1.1-S1.4 to Become Organization Unfolding Structure

The model resulting from steps S1.2-S1.4 is then used to produce what is known as the unfolding structure of the MyFlix organization/system (Fig. 7). The purpose of this model is to look at the recursive structure inherent in the organization that aims to balance the load. It is much different from the structure as it appears in the hierarchical organizational chart to show unstable power (higher upwards more powerful). The structure of the unfolding structure of the MyFlix organization/system resulting in an overview of the organizational structure is not based on the power but on how an organization is formed based on the elements identified in steps S1.1 to S1.4. In the figure it can also be seen how MyFlix organizations are scaled up in the early stages based on technology, geography, client and suppliers as well as time. This makes the design of MyFlix application scalable.

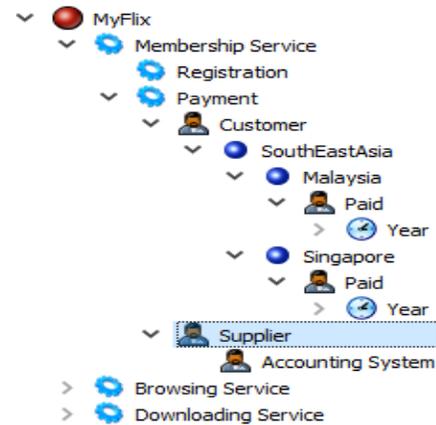


Fig. 7. MyFlix unfolding structure.

G. S2: Model the Business Process

This step is aimed at identifying the business processes involved in each branch of technology. This step can be done by referring to the existing business process or creating a new business process and matched to a recognized branch of technology. Here is the identified business process for the MyFlix technology branches. Step S2 consists of sub steps starting from S2.1 to S2.2.

H. S2.1 for each Branch from the Root of the Organization's Unfolding Structure Resulting from S1.5

In this step, the level of technology of one Membership Service was selected for modelling the business process. Modelling other branches such as Browsing Services and Download Services will be implemented upon completion of the Membership Services branch. It can be run simultaneously by other developer groups. There is no branch sequence that needs to be modelled first.

I. S2.1.1 Start from the Bottom Level Technology Branch N to Level 0

The second level is the bottom level. This level is comprised of a branch of Payment and Payment technology. Then followed by the one level of Membership Services.

J. S2.1.1.1 for each Level, Model the Business Process for each Branch of Technology at that Level

K. S2.1.1.1.1 if the Level Is the Leaf Level, Model the Business Process by Taking into Account All Types of Complexity Drive that Will Use it at the Higher Level

Since the second level is the leaf level then modeling the business process is done on the technology branch of Registration (Fig. 8) followed by Payment (Fig. 9).

L. S2.1.1.1.2 if it is not a Leaf Level, Model the Business Process by Taking into Account All Types of Complexity Drive that will Use it at the Higher Level and All Types of Complexity Drive used at the Lower Level

After completion of the leaf level, the next level is the level of Membership Services (not the leaf level) (Fig. 10). The Business Services business process needs to take into account the business processes of Registration and Payments before modelling it. It also needs to take into account the upper branch of MyFlix. At MyFlix level it involves clients to organizations.



Fig. 8. Registration business processes.

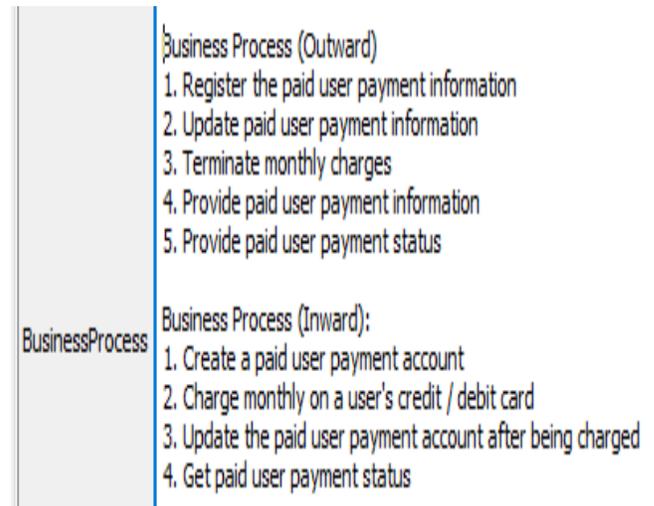


Fig. 9. Payment business processes.

M. S2.2 Convert the Resulting Business Process into the Form of Promise Statement

Once completed modelling the business process for each focused level, the next step is to convert the business process into a promise form. For each branch of MyFlix technology, promise statements are generated for Membership (Fig. 11), Payment (Fig. 12) and Registration (Fig. 13) services. The purpose of the promise statements are to be modelled so that it can convert a coercive business processes to a more voluntary form based on the Promise Theory.

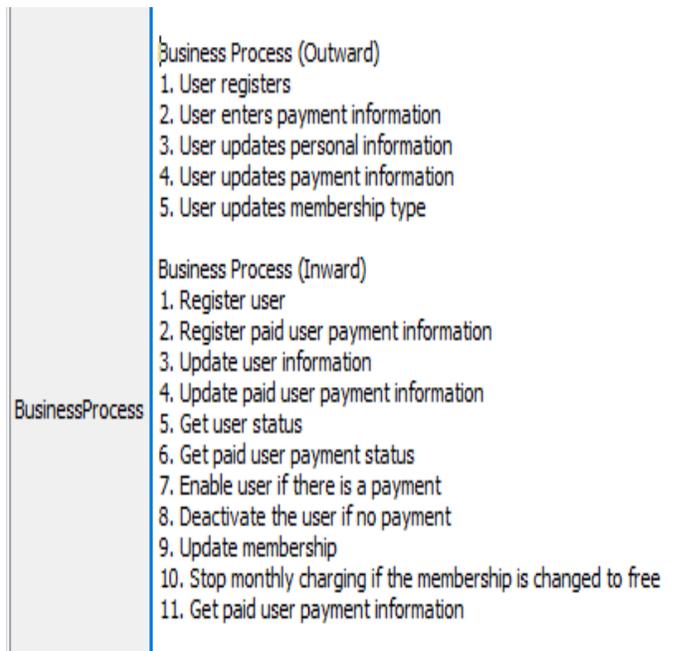


Fig. 10. Membership business processes.

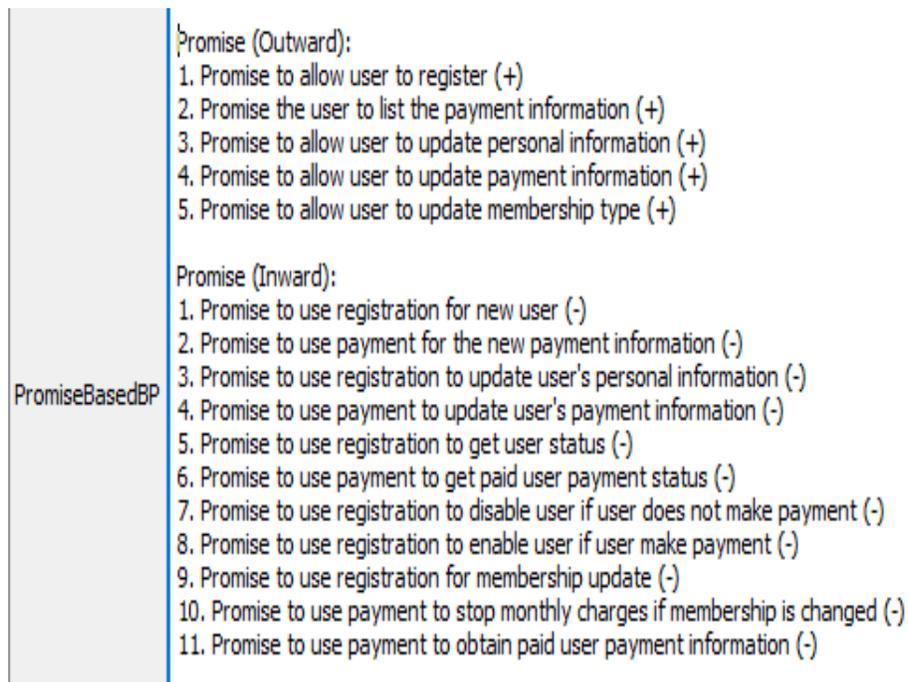


Fig. 11. Membership promise statements.

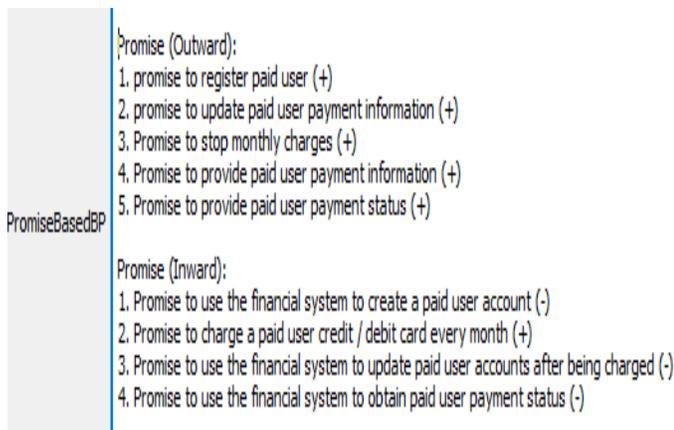


Fig. 12. Payment promise statements.

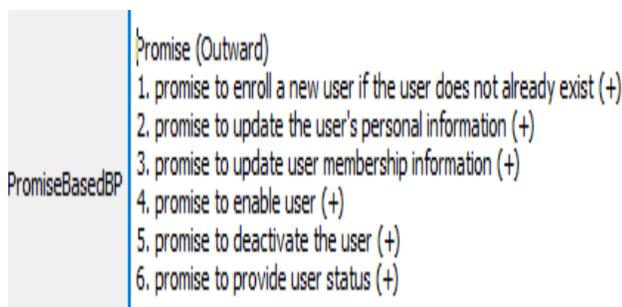


Fig. 13. Registration promise statements.

N. S2.3 Model the Promise Diagram based on Promise Statements from Step S2.2

The promise statement resulting from step S2.2 is used to model the promise diagram. This diagram is very useful to see the balance of interactions between agents (technology branches). The promise theory emphasizes equilibrium and dynamic aspects of equilibrium. It also emphasizes the convergence of interacting agents that are important to produce a stable system. The theoretical model of the promise that has been made between the Membership Service with Registration and Payment (Fig. 14). This is to indicate that there is an interaction between Membership Services with Registration and also what promises are exchanged between the two agents. The promise statement must have a symbol (+) meaning the promise given while (-) means a promise of use. It can also be seen in the figure of the interaction between Membership Services and Payments.

O. S3: Generate Microservice

Step S3 aims to produce microservice after the completion of the S2 step is implemented. The microservices to be produced are services that have APIs. The resulting microservices are general and can be implemented in any microservice framework framework such as Spring Boot.

P. S3.1 List the Microservice from Step S2.3

Microservice candidates will be taken from the model diagram produced in S2. Agencies that promise and use promise are eligible as microservice candidates. In the case of MyFlix, for the technology branch of the Membership Service, microservice are:

- 1) Membership service
- 2) Registration, and
- 3) Payment.

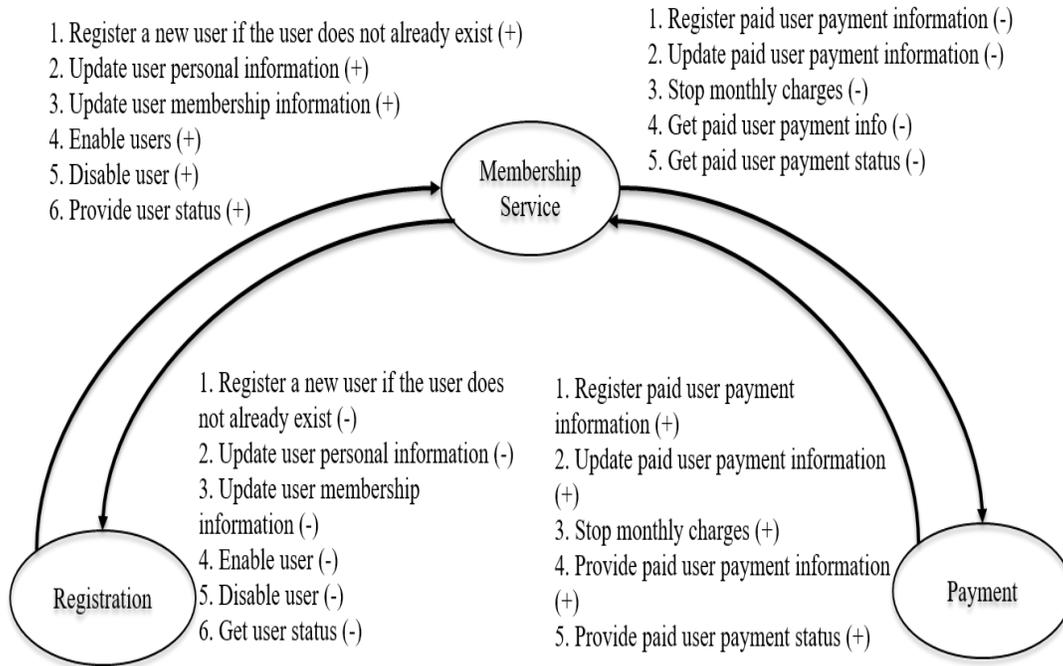


Fig. 14. Promise diagram of membership service.

Q. S3.1.1 Determine the API for each Microservice
Candidate based on an Promise Model from Step S2.2

After the candidate microservice are identified in Step S3.1 then the next step is to determine the API for the candidates of microservice. Referring to the model of a promise such as API diagram can be named based on the promises of one agent to another agent. For example, Registration promises to the Membership Service to “register a new user if the user does not already exist (+)” then the API for the Registration microservice can be named with “registerUser()” but still refers to the same promise by Registration to Membership Service. The resulting APIs from S3.1.1 step for Membership (Fig. 15), Payment (Fig. 16) and Registration (Fig. 17) services.

```

Microservice
public register ()
public registerPayment ()
public updatePersonalInfo ()
public updatePayment ()
public updateMembershipType ()

private registerUser ()
private registerPaymentInformation ()
private updatePersonalInfo ()
private updateInformationPayment ()
private userStatus ()
private paymentStatus ()
private disableUser ()
private activateUsers ()
private updateMembership ()
private stopCharging ()
private userPaymentInformation ()
    
```

Fig. 15. Membership microservice API

```

Microservice
public registerPaymentInfo ()
public updatePaymentInfo ()
public stopCharging ()
public paymentInfo ()
public paymentStatus ()

private create account ()
private charge ()
private updatePayment ()
private userPaymentStatus ()
    
```

Fig. 16. Payment microservice API.

```

Microservice
public registerUser ()
public updateUser ()
public updateMembership ()
public activateUser ()
public deactivateUser ()
public userStatus ()
    
```

Fig. 17. Registration microservice API.

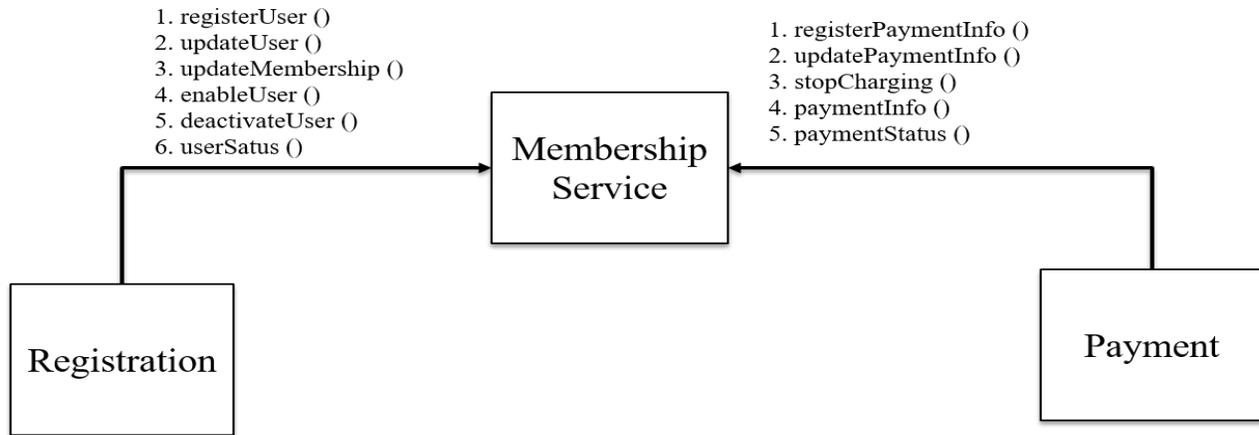


Fig. 18. Membership dependency diagram.



Fig. 19. Aligning organization to business processes and microservices.

R. S3.2 Model a Microservice Dependency Diagram

It is possible to see how a microservice depends on each other (Fig. 18). In microservice architecture, the two main types of interactions that are the first are synchronous and the second is asynchronous. Examples of synced interactions are the use of keywords such as Get, Post, Put and Delete which are the dialect used for REST interactions. Examples of asynchronous interactions are by using technologies such as pub/sub. It is up to the builder to implement microservices either asynchronously or synchronously.

S. S4: Map the Organization's Unfolding Structure to the Business Process to Microservices

The last step is to map the unfolding structure of the MyFlix organization/system to the business process and to the microservice that have resulted from the S1-S3 step. This step is important as it will align between the organization, the business and the microservice. The result is a fully-documented MyFlix-based application structure and ready to be updated in the next iteration. It can also be considered as a knowledge base as well as a map to MyFlix applications which can be consulted and understood by developers anytime and anywhere MyFlix application developers are located. The mapping of the MyFlix unfolding structure with business processes and microservices for the branch of Membership (Fig. 19).

VI. RESULTS

Experts have been interviewed to give expert opinion on the methods that have been proposed. The experts background (Table I).

The number of experts is only 5 because of the difficulty to get experts who are familiar with microservice architecture in Malaysia. However, even though only 5 experts have evaluated the method, the experts chosen represent developers from the industrial and also from the government sectors that has long experience in developing applications. Before the interviewing process, the experts have gone through the demonstration of the method of the case study in Section V. The process is called a walkthrough process. Then, interview has been conducted to get as a way to evaluate the proposed method and to get the expert opinion. Based from the interview with the experts, the results gathered from the interview are as follows:

TABLE I. EXPERTS BACKGROUND

Expert	Sector	Experience as Software Developer
1	E-Business	14 years
2	Software House	15 years
3	Government	13 years
4	Government	13 years
5	Government	11 years

- 1) All experts agree that this method would be able to design microservice-based application systematically.
- 2) All experts agree that this method would be able to reduce the complexity of developing distributed system.
- 3) All experts agree that this method would be able to design a stable microservice-based application.
- 4) All experts agree that this method would be able to effectively document the design process.
- 5) All experts agree that this method would be able to align the organization with the business and the information technology.
- 6) All experts agree that this method would be able to design a scalable microservice-based application.

Among key insights gathered from the interview with the experts are as follows:

- 1) The method can be used by developers in the industrial and government sectors for transforming existing monolithic application that are not scalable and expensive to maintain.
- 2) In developing complex system, this method is important to help developers to first design before rushing into the implementation as the latter can cause system instability and spaghetti system and the end result is a monolithic application.
- 3) The method not just help to align the IT to business but most importantly aligning to the organization structure. Therefore, it is important to first design a distributed and decentralized organization structure that confirm will produce a loosely coupled but cohesive business processes and microservices.

VII. FUTURE RESEARCH

The artefacts produced in this study can be used and tested for their efficacy in diverse case studies. In this study, this solution method is tested on a case study that provides online service-based applications constrained by monolithic architecture. The outcome of a case study in this study can be a benchmark for other studies that use different cases from the case in this study.

The method that has been developed from this study is specifically to produce a microservice-based application design that can be implemented in microservice architecture. However, this method is also likely to be modified for use in research involving the creation of any distributed and decentralized system design. For example, studies in Internet-of-Things (IOT) involve any tools such as hardware and sensors that need to be designed so that they can achieve their intended purpose. The study in this area is still new and the solution method from this study is seen to have potential for use in the IOT field. Apart from the IOT field, any area requiring distributed and decentralized system designs can also take advantage of this method as in cyber security, blockchain and business modeling.

VIII. CONCLUSIONS

This study is an attempt to produce an artefact in the form of a method that can be used to design scalable microservice-based application systematically. Starting with the problem of lack of methods to produce scalable microservice-based application designs it has been suggested in this study to find a method that can be used to produce scalable microservice – based application design. The result findings have shown the method capabilities to be used to model a scalable microservice-based application design. The principles behind the method has made the proposed method to be able to deliver not just scalable design of microservice-based application but also taking into consideration the stability of the application that will be produced. However, the method produced from this research is limited only for designing a scalable microservice-based application without considering other aspects such as cost, security and human.

The contribution of this method is very important as existing methods are still not systematic and do not guarantee the development of scalable microservice-based applications. This method is designed specifically for developers to design microservice-based applications that will guarantee the development of scalable microservice-based applications. It is also important because it can align between organization, business and information technology for the sake of organization viability.

REFERENCES

- [1] Beer, S. 1979. The heart of enterprise. Managerial cybernetics of organization. Wiley.
- [2] Bhadoria, R. S., Chaudhari, N. S. and Tomar, G. S. 2017. The Performance Metric for Enterprise Service Bus (ESB) in SOA system: Theoretical underpinnings and empirical illustrations for information processing. *Information Systems* 65: 158–171.
- [3] Burgess, M. 2015a. Thinking in Promises: Designing Systems for Cooperation. O'Reilly Media.
- [4] Burgess, M. 2015b. In search of certainty: the science of our information infrastructure.
- [5] Conant, R. C. and Ross Ashby, W. 1970. Every good regulator of a system must be a model of that system. *International journal of systems science* 1(2): 89–97.
- [6] Conway, M. E. 1968. How do committees invent. *Datamation* 14(4): 28–31.
- [7] Espejo, R. and Reyes, A. 2011. Organizational Systems: Managing Complexity with the Viable System Model. Springer Berlin Heidelberg.
- [8] Evans, E. 2013. Domain-Driven Design Quickly. *Journal of Chemical Information and Modeling*, p. Vol. 53.
- [9] Friedrichsen, U. 2017. Resilient Functional Service Design. InfoQ.
- [10] Gucer, V., Narain, S. and others. 2015. Creating Applications in Bluemix Using the Microservices Approach. IBM Redbooks.
- [11] Gysel, M., Kölbener, L., Giersche, W. and Zimmermann, O. 2016. Service Cutter: A Systematic Approach to Service Decomposition. *European Conference on Service-Oriented and Cloud Computing*, p. 185–200.
- [12] Levcovitz, A., Terra, R. and Valente, M. T. 2016. Towards a Technique for Extracting Microservices from Monolithic Enterprise Systems.
- [13] Newman, S. 2015. Building Microservices. O'Reilly.
- [14] Pahl, G., Wallace, K., Blessing, L. T. M., Beitz, W. and Bauert, F. 2013. Engineering Design: A Systematic Approach. Springer London.
- [15] Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*.
- [16] RV, R. 2016. Spring Microservices. Packt Publishing.
- [17] Timms, S. 2016. Mastering JavaScript design patterns. Packt Publishing Ltd.
- [18] Vernon, V. 2013. Implementing Domain-Driven Design. Pearson Education.
- [19] Yahia, E. B. H., Réveillère, L., Bromberg, Y.-D., Chevalier, R. and Cadot, A. 2016. Medley: An Event-Driven Lightweight Platform for Service Composition. *International Conference on Web Engineering*, p. 3–20.

Adaptive Simulated Evolution based Approach for Cluster Optimization in Wireless Sensor Networks

Abdulaziz Alsayyari

Computer Engineering Department
Shaqra University
Dawadmi, Ar Riyadh, Saudi Arabia

Abstract—Energy consumption minimization is crucial for the constrained sensors in wireless sensor networks (WSNs). Partitioning WSNs into optimal set of clusters is a promising technique utilized to minimize energy consumption and to increase the lifetime of the network. However, optimizing the network into optimal set of clusters is a non-polynomial (NP) hard problem, and the time needed to solve such problem increases exponentially as the number of sensors increases. In this paper, simulated evolution (SimE) algorithm is engineered to tackle the problem of cluster optimization in WSNs. A goodness measure is developed to measure the accuracy of assigning nodes to clusters and to evaluate the clustering quality of the overall network. SimE was developed such that the number of clusters and cluster heads are adaptive to number of alive nodes in the network. In fact, extensive simulation results demonstrate that SimE provides near optimal clustering and improves the lifetime of the network by about 21% compared to the traditional LEACH-C protocol.

Keywords—Clustering algorithm; cluster optimization; network lifetime; simulated evolution; wireless sensor networks

I. INTRODUCTION

Wireless sensor networks (WSNs) are formed using small sensor nodes to monitor certain phenomena of environments where human presence may be impossible or not preferred. After wireless nodes are deployed and connected together, data about sensed events is typically gathered and reported to a centralized location for further processing [1], [2]. Nevertheless, the applications of WSNs are wide and vary from one application to another [3]. The application often customizes the details of designing wireless sensor nodes and WSNs' planning; including node architecture, communication protocols, network topology, and deployment schemes [4]. In large-scale deployment scenarios of WSN such as battlefields and forest habitat monitoring, sensor nodes often have limited resources. This is because batteries in such deployment scenarios are mostly neither changeable nor chargeable. As a result, batteries of sensor nodes are considered a sacred resource [5]. Therefore, minimizing energy consumption is necessary to increase the life time of the WSN.

In WSNs, data is exchanged between sensor nodes in an ad hoc fashion. This technique allows the network to cover larger geographical areas, extend the reach of the network, and help sensor node in saving energy by lowering transmission power of the node and allowing neighboring nodes to perform certain network duties alternately [6]. In fact, clustering is a popular method commonly utilized in WSN to prolong network

lifetime [7]. Furthermore, efficient clustering directly leads to energy saving and, hence, results in extending network lifetime [8], [9]. Clustering is achieved by grouping a specific set of sensor nodes in one cluster and, then, assigning a cluster head (CH) to handle certain tasks in the cluster. Typically, nodes are selected in one cluster according to criteria such as cluster size and nodes' locations. In such scenario, nodes in the cluster communicate with the cluster head instead of directly communicating with the base station (BS). Later, the cluster head aggregates packets received from cluster nodes and sends them to a BS.

In this work, a simulated evolution (SimE) algorithm for cluster optimization in WSNs to provide near optimal solutions is presented. More specifically,

- A simulated evolution (SimE) algorithm is developed to cluster the WSN and increase its lifetime. The results show that the proposed SimE algorithm minimizes energy consumption in the network. This is achieved by minimizing the total sum of squared distances between cluster nodes and the CHs.
- A goodness measure, which is the core part of the SimE algorithm, is proposed to tackle the WSN clustering problem and to evaluate the quality of the produced clusters.
- Unlike previous methods, the adaptivity of number of clusters (or CHs) to the network size is also addressed, where it is shown that number of clusters is adaptive to number of alive nodes in the network and a clustering algorithm should be adaptive to number of alive nodes per round instead of assuming a fixed number of clusters. The adaptivity of the proposed SimE approach eliminates the need to develop a multi-objective optimization function to account for load balancing of the clusters.
- This paper investigates the effect of BS location and deployment area on the network lifetime and addresses the change in number of clusters as deployment area changes.
- The simulation results show that the proposed SimE approach enhances the network lifetime by about 21% compared to the LEACH-C protocol.

The remainder of this paper is organized in the following order. While Section II provides a background study on the

research literature in relevance to cluster optimization in WSNs, Section III presents an overview of the proposed SimE method, including the algorithm and the goodness measure details. Meanwhile, Section IV provides the performance results and finding of the proposed SimE approach. Finally, Section V concludes the work.

II. RELATED WORK

K-means algorithm is a popular approach utilized in WSNs among many other applications to produce clusters [10], [11]. In fact, various approaches that are based on such algorithm have been developed to ensure a more efficient clustering [12], [13]. Such development efforts are continuously attempted as a result of inherent challenges that exist in the WSN clustering scheme. The objective of this problem is to find k optimal clusters such that the total energy is minimized and the lifetime of the WSN is increased. The nodes of the network are grouped (clustered), where they are either member nodes or CH node. Member nodes send to the CH instead of sending directly to the BS. This allows for a reduction in the communication distance and an increase of the lifetime of the network [8]. In general, the number of clusters and CHs are not previously known. Therefore, this number might change over time due to the complete energy loss in some nodes in the network, which further complicates the problem of finding optimal clustering using k-means algorithm. In fact, the problem of k optimal cluster optimization in WSNs was proven to be non-deterministic polynomial-time hard (NP-hard) problem [14].

Many evolutionary approaches and protocols targeting cluster optimization were proposed for WSNs. One of the most well-known approaches is the Low Energy Adaptive Clustering Hierarchy (LEACH) protocol, which is a distributed clustering algorithm [15]. In the LEACH protocol, CHs are randomly selected. Then, they advertise their presence by utilizing the Carrier Sense Multiple Access (CSMA), which is a Medium Access Control (MAC) protocol. Cluster members (CMs) that have not been selected as CHs choose the corresponding CH based on the received signal strength (RSS). In fact, they send their packets to the corresponding CH instead of sending them to the base station (BS) to reduce the energy consumption of the CMs. The CH, however, aggregates the received packets into a single message and forwards it to the BS using spreading codes and CSMA/MAC protocol.

It was shown in [8] that random selection of CHs using decentralized approaches as in LEACH is not efficient in terms of energy consumption. It was also shown that using centralized approach increase the lifetime of the network since it is possible to rotate the selection of the CHs in each round. Furthermore, number of CHs is proportional to the network energy consumption, which directly affects the network lifetime [16]. In fact, the study in [8] proposed a revised version of LEACH called centralized LEACH (LEACH-C) protocol. Generally, the BS centrally configures the clusters according to the communication distance and the energy levels obtained from network's nodes. In LEACH-C protocol, however, the simulated annealing (SA) algorithm is utilized to configure clusters [17]. To balance the energy consumption in each node, only nodes that have energy levels greater than the

average energy of the nodes are nominated to be CHs. The BS runs SA to form the clusters by utilizing the nominated CHs.

A genetic based approach was proposed in [18] and several factors affecting the optimization of the clusters such as the BS location were discussed. In fact, the number of clusters is not adaptive, which may cause uneven number of nodes in the clusters, and it thus was assumed that number of CHs is 10%. The results in [18], however, showed that as number of nodes doubles, the population size needs to be doubled as well for the purpose of maintaining comparable performance. 80% reduction in the distance, on average, was achieved compared to the distance obtained by direct transmission. Several studies discussed QoS routing in WSNs including [19], [20], [21]. The study in [19] presented a multi-objective genetic algorithm for efficient QoS routing in two tiered WSNs. Three fitness functions were introduced to form the multi-objective function of the genetic algorithm, which are energy consumption, delay and reliability. Additionally, it was shown that genetic algorithm is reliable in optimizing these functions including QoS in WSNs. However, performance results of the proposed genetic algorithm in terms of number of alive nodes per round were not reported.

Nevertheless, several studies utilized the particle swarm optimization (PSO) algorithm for cluster optimization in WSNs; for example, see [16] and [22]. In [22], however, the objective function is formed from the Euclidean distance of nodes and the energy consumption of nodes in each round. A constant was utilized to weigh these functions to form the multi-objective function. The proposed PSO approach was compared to LEACH-C to demonstrate the effectiveness of PSO algorithm. Though, unequal initial energy for the nodes and a fixed (not adaptive) number of CHs were assumed. Furthermore, the study in [16] utilized PSO for cluster optimization in WSNs. In this study, the PSO algorithm aims to minimize energy consumption by minimizing number of active CHs. However, minimizing CHs is not always the most appropriate strategy for minimizing the energy consumption as was explained in [8].

Furthermore, the study presented in [23] provided WSNs clustering algorithms based on simulated annealing (SA) and PSO algorithms. Their approach was presented to provide better clustering when compared to LEACH protocol. However, their objective function is actually a multi-objective, which allows less flexibility for energy load distribution among the clusters and that number of clusters in their approaches is fixed. Meanwhile, Tabu search based centralized approach was proposed in [24] for cluster optimization in WSNs. The nodes and the connection between them were represented as a hypergraph, which is a graph with edges having multiple nodes. This approach initially represents the cluster nodes and their CH as a Clique and apply Tabu search to optimize the Clique problem. Although the authors showed that their proposed Tabu search outperforms SA algorithm, the runtime of their proposed approach is higher than SA and it requires addressing many complicated Tabu search related structures such as short, medium, and long-term memories. Recently, the new nature inspired Cuckoo search algorithm was applied for cluster optimization in WSNs, for example, see [25]. In this study, the approaches aims to optimize randomly created

clusters. However, no description of how the Cuckoo search was applied to the clustering optimization, which is discrete in nature, giving that cuckoo search is mainly developed for continuous objective functions.

Nevertheless, many other proposals attempted to solve the problem of clustering the WSN as a pure clustering problem, for example, see [14], [26] and [27]. For this to work, fixed cluster need to be assumed, which is not suitable for the problem of clustering WSNs due to the change in number of nodes in the network over time. In addition, pure clustering might perform worst when the number of nodes decreases in the network due to the complete loss of energy.

III. DESCRIPTION OF THE PROPOSED APPROACH

In this section, a full description of the proposed SimE approach is provided, which is being utilized to optimize number of clusters in the network. This includes describing the SimE algorithm, the proposed goodness measure, and how SimE is engineered for cluster optimization in WSNs.

A. Assumptions

In this work, the following are assumed (no assumption about network density is made):

- The BS has unconstrained power source.
- Each sensor node belongs to exactly one cluster.
- The sensor nodes are static given that in the majority of applications sensor nodes have no mobility.
- Initially all sensor nodes are charged with the same amount of energy.
- Communication links are bidirectional.
- The computation and communication capabilities are the same for all network nodes.
- The only source of energy in sensor nodes is the battery.
- The sensor nodes are unaware of their location. Most of the contributions found in the literature assumed that the sensors can determine their location by means of the Global Positioning System (GPS), which is an unrealistic assumption. This paper, however, adopts the approach described in [16], which assumes that each sensor maintains a list of its neighbors. In that work, the flooding method is utilized to send the list to the BS, where it can decide which nodes will be CHs based on the information received.

B. SimE Description

SimE is a very attractive and elegant evolutionary iterative algorithm that is being utilized over the years to solve various types of optimization problems. By employing SimE, the search space is traversed in a smarter way using smart moves, which makes it outperform other iterative algorithms for most different problems. The evolution of the SimE is as follows: first ill-assigned nodes are determined, and they become candidates for moving them to a better cluster. With iterations, the quality of the solution is improved as the ill-assigned nodes either decrease in number or placed in the best possible cluster.

Therefore, unlike other iterative algorithms such as genetic algorithms, SA and PSO, the evolution of SimE with iterations is smarter and more efficient.

Typically, the SimE algorithm consists of three main steps that are executed in sequence; the evaluation, selection and allocation steps as described in Fig. 1 [28]. In the evaluation step, the nodes are evaluated based on the goodness of each node in the cluster solution and ill-assigned nodes are marked to be considered for movement. Note that in order for the SimE to escape local minima, some nodes that are good might be chosen based on some random parameter. The selection step performs this and also puts nodes to be moved in a selection list PS as in Fig. 1. The allocation step allocates the selected nodes to the clusters based on checking best cluster of the current solution.

In [28], the SimE was selected for cluster optimization in WSNs because it is believed that it is naturally more suitable for cluster optimization in WSNs. This is believed because of the following reasons. Firstly, the nature of WSNs clustering depends on choosing CHs, which are not necessarily the same each round. And, there are certain nodes that can join/leave certain cluster at certain round. Secondly, nodes might die or completely lose energy including CHs, which is not a problem for SimE, given that these nodes simply can be discarded from the clusters without affecting other clusters. This also suggests that the SimE is adaptive in this regard. For example, number of clusters could be reduced at certain round due to the complete loss of energy of all its nodes. Unlike other protocols and heuristics, the SimE itself determines, given an upper bound, the best number of clusters at each round.

```
ALGORITHM Simulated Evolution(N,  $\emptyset_{initial}$ );
/* B: Selection bias; */
/*  $n_i$ : node i; */
/*  $g_i$ : goodness of node  $n_i$ ; */
/*  $\emptyset$ : Complete Solution; */
INITIALIZATION;
Repeat
EVALUATION:
ForEach  $n_i \in \emptyset$  Do Evaluate  $g_i$ , EndForEach;
SELECTION:
ForEach  $n_i \in \emptyset$  Do
If Selection( $n_i, B$ ) Then PS = PS  $\cup$  { $n_i$ }
EndIf;
EndForEach;
Sort the elements of PS;
ALLOCATION:
ForEach  $n_i \in PS$  Do Allocation( $n_i, \emptyset$ ) EndForEach;
Until Stopping-criteria are met;
Return (BestSolution);
End Simulated Evolution.
```

Fig. 1. Simulated evolution structure [25].

In general, the main component of the algorithm is the goodness measure. Such measure requires to be carefully developed in order to get a good quality final solution produced by the SimE. The goodness value indicates how well a certain cluster node is currently assigned. In such case, the

higher the value of the goodness provided, the lower the probability of the node being selected for reallocation is. In fact, allocation is the most important step in SimE algorithm and has the most impact on the quality of the produced solution. The selection set PS and the partial solution \emptyset_i are the inputs of the allocation operator. A new complete solution \emptyset_N is generated according to an allocation function, which depends on the optimization problem being solved and generally allocates the elements in the PS. The importance of the allocation step comes from the iterative improvement, where previous solution is improved as PS elements are being assigned to a better cluster, without being too greedy.

C. Goodness Evaluation

The idea of the presented goodness measure is to utilize the fact that a node is considered for moving it from the current cluster to another cluster if its goodness value in the current cluster is low. To determine the goodness of a node in a cluster, one must find the total cost of the cluster nodes when a direct communication is made between them and BS. Then, a calculation should be performed of the total cost when one of the cluster nodes is randomly selected as a CH. This represents the goodness of the cluster. To find the goodness of a node in the cluster, the distance from the node to the nodes in the cluster is divided by the total cost from the cluster nodes to the BS. The lower the goodness value of the node is, the higher probability it is to move the node from the cluster to another cluster. To illustrate this, consider the example in Fig. 2 for one cluster having four nodes and a BS, assuming node c is the CH.

In the above case, the goodness of node a (gda) will be:

$$gda = 1 - ((d0 + d1 + d2 + d8) / (d6 + d7 + d8 + d9)).$$
 Similarly,

$$gdb = 1 - ((d1 + d3 + d5 + d8) / (d6 + d7 + d8 + d9))$$

$$gdc = 1 - ((d2 + d3 + d4 + d8) / (d6 + d7 + d8 + d9))$$

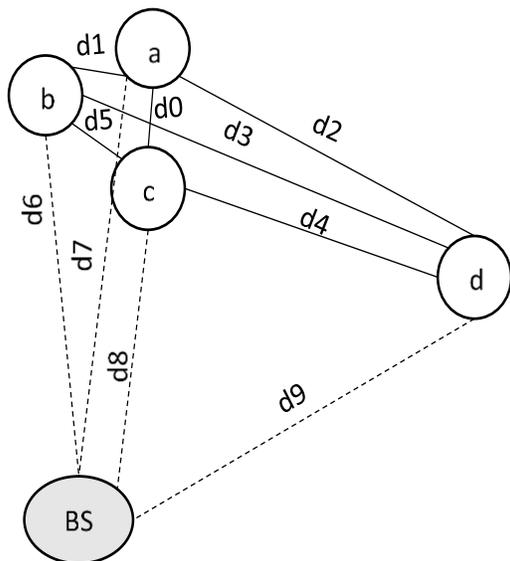


Fig. 2. Illustration of the goodness measure.

The goodness gdc is less than gda and gdb and, hence, node d is less likely to be part of the cluster and will be considered for movement to another cluster. However, node d

might not move into another cluster if gdc is the best possible goodness for all clusters. This illustrates the need for making number of clusters adaptive to maximize the lifetime of the network. The goodness of the assumed CH c is calculated in the same way and might move to another cluster, where another cluster node will act as a CH. Therefore, CH selection is not important in the presented algorithm as any one of the cluster nodes can act as a CH with some very little increased energy, which will be discussed in more details in the results section.

IV. SIMULATION RESULTS AND ANALYSIS

This section explains the simulation implementation, illustrates the results, and provides an analysis of the obtained results.

A. SimE Performance

SimE was implemented in C++ and a random initial network was created. It is worthwhile to mention that the BS is responsible about running the SimE and producing the clusters. Hence, no extra work is needed by the CHs except the aggregation of the data collected from the CMs. Nodes were also deployed randomly. For the simulation, a laptop with Intel i5 processor, 8G memory, and 750G of physical memory was utilized. To demonstrate the output of the presented SimE implementation, Fig. 3 shows an example of clustering 100 nodes in an area of 50X50 m². The BS was placed at (50,175). The upper bound of number of clusters was set to 100 and number in the figure represents the cluster that the node in that location is assigned to.

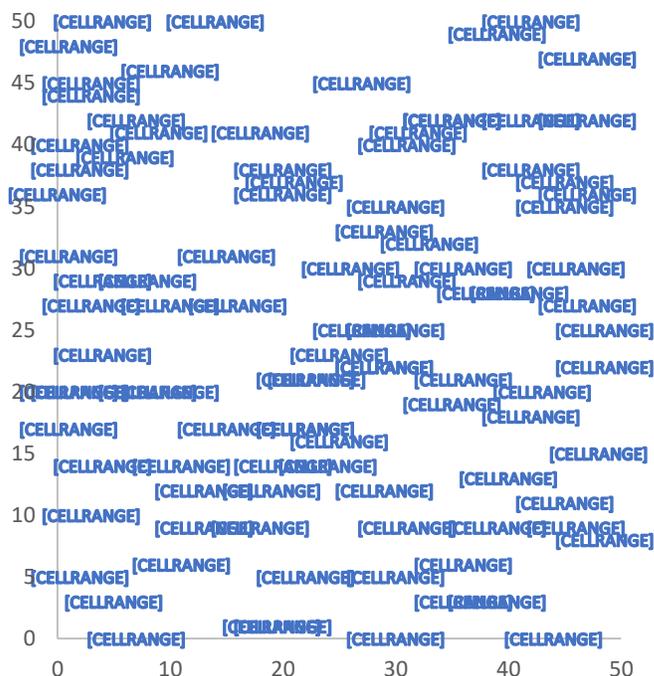


Fig. 3. The resulting clustering of SimE for 100 nodes in a 50X50 m² area.

After many experiments, B value was determined to be 0.1 and a balance parameter was introduced to balance the energy load among the clusters. If a node is considered for allocation

in a cluster x and number of nodes in x exceeded the limit, i.e., $(\# \text{ of nodes in the network} / \# \text{ of clusters}) + \text{balance}$, the allocation of the node to the cluster x will be discarded and the node will be allocated to another cluster.

To evaluate the quality of the solution produced by the SimE and to test the goodness measure being proposed, many experiments were carried out. Considering a network deployed in a $100 \times 100 \text{ m}^2$ area and a BS located at $(50, 175)$, Fig. 4 depicts the overall average goodness over 1000 iterations. The simulation is repeated 100 times and the average goodness was taken. As can be seen from the figure, the goodness initially starts at low value which indicates that the initially random clustering is not good and start improving till it reaches above 0.9 and starts the hill climbing process trying to escape local minima. The maximum average goodness is 0.97. As expected, selection list PS is behaving the opposite of the goodness and is generally decreasing through iterations as in Fig. 5.

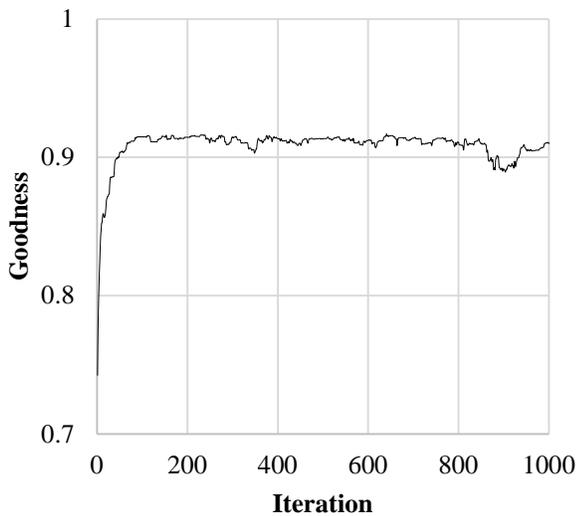


Fig. 4. The behavior of the average goodness over iterations.

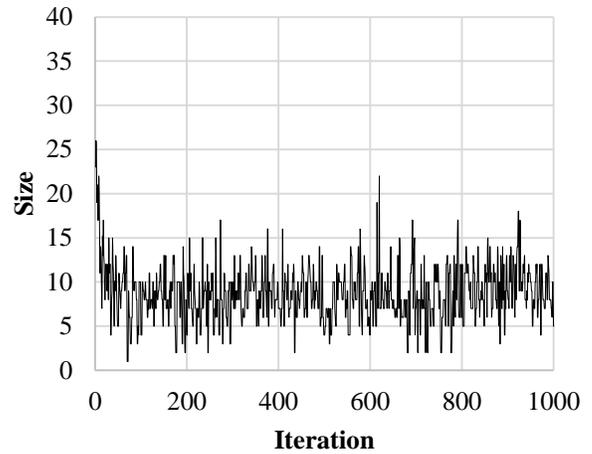


Fig. 5. Selection list size over iterations.

Table I summarizes the results obtained by the SimE for an area of $100 \times 100 \text{ m}^2$. The table shows the goodness, number of iterations taken till no improvement in the solution, the percentage of reduction in the distance compared to direct communication between the nodes and the BS, the runtime and number of clusters produced. The table was produced based on the average of 100 runs. The goodness increases when number of nodes increases and also the reduction in the distance tends to increase as number of nodes increases. These results suggest that the upper bound of number of clusters should be higher when number of nodes increases because the cluster coefficient will be higher in smaller area. The results also demonstrate that the location of the BS plays an important role in the optimization of the clusters. This is due to the fact that the distance of some nodes from the BS is less than the distance to any other node in the network and, therefore, it is concluded that when optimizing and forming clusters, the BS should be far from the area of the deployment to obtain good clustering.

Further, to study the relationship between the cluster optimization and the area of deployment, simulations for an area of $200 \times 200 \text{ m}^2$ were also carried out. The simulation results of this experiment are shown in Table II.

TABLE I. SUMMARY OF PERFORMANCE FOR AREA OF $100 \times 100 \text{ M}^2$

Area	100X100											
BS location	500,500				200,200				100,100			
Node size	100	200	400	800	100	200	400	800	100	200	400	800
Goodness	.98	.98	.98	.98	.95	.95	.95	.94	.79	.74	.85	.87
Iterations	625	766	771	952	330	506	847	960	22	26	33	99
Distance reduction	81.3	88.2	93	95.3	78.5	84.5	88.4	90.8	71.2	73.1	77.9	80
Time (sec)	.39	.71	2.6	7.74	.34	0.64	3.86	5.85	.07	.15	.2	.45
# clusters	17	20	20	20	16	20	20	20	14	18	40	77

TABLE II. SUMMARY OF PERFORMANCE FOR AREA OF 200X200 M²

Area	200X200											
BS location	500,500				400,400				300,300			
Node size	100	200	400	800	100	200	400	800	100	200	400	800
Goodness	.93	.94	.96	0.97	.93	.95	.96	.97	.92	.92	.94	.96
Iterations	174	226	314	400	148	217	321	409	166	169	296	366
Distance reduction	86.4	88.1	90.4	92	82.98	84.9	86.7	88.1	75.8	82.9	85.2	86.4
Time (sec)	.15	1.55	.94	3.35	.12	.61	2.71	6.14	.11	.32	1.5	7.81
# clusters	5	10	20	40	10	18	38	75	18	19	38	78

Since the area size is 200X200 m², the BS was placed outside the deployment area by selecting 300X300 m² and 400X400 m² locations. There is a fluctuation in the distance reduction and the goodness since the network is deployed randomly. Furthermore, no assumptions were made regarding the clustering coefficient, in which the nodes tend to be groped to some extent while randomly deploying the network. For this reason, clustering coefficient will be lower for larger deployment area. Therefore, it is better to increase the upper bound of number of clusters for larger area.

Fig. 6, however, depicts the relationship between the deployment area, the distance reduction, and number of clusters for 100 nodes and BS located at (250,250). As the deployment area increases, number of clusters decreases and the distance reduction increases until some point. The number of clusters decreases as the area increases. This finding reflects the fact that the number of clusters should not be assumed fixed at each round as most of contributions in the literature assumed, i.e., 5%, while it needs to be adaptive in order to account for the deployment area size. Moreover, as the deployment area increases, the parentage of distance reduction tends to decrease. The reason of this is that the distances between the CHS and the BS increases, which does not mean a low quality clustering. On the other hand, number of clusters is also affected by the location of the BS, which also demonstrates the need for making number of clusters adaptive for better clustering results.

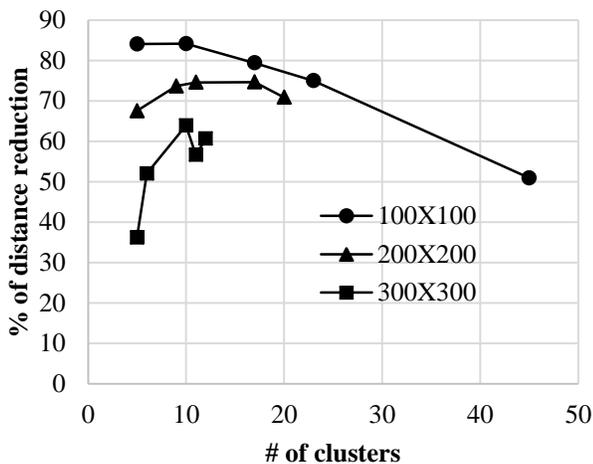


Fig. 6. The effect of the area size on the percentage of the distance reduction and number of clusters.

B. Network Lifetime

To investigate the network lifetime using the proposed SimE approach, experiments were also conducted in C++. The energy consumption model is assumed to be the same as in [8] and [29]. The energy consumed to transmit (ETx) and receive (ERx) 1 bits of packet over a distance d in radio hardware can be written as in (1) and (2), respectively.

$$ETx(l, d) = \begin{cases} 1 Eelc + l efs d^2, & d < d_0 \\ 1 Eelc + l emp d^4, & d \geq d_0 \end{cases} \quad (1)$$

$$ERx(l) = 1 Eelc \quad (2)$$

$$d_0 = \sqrt{\frac{efs}{emp}} \quad (3)$$

Where efs and emp are factors of energy dissipation rate in the power amplifier and Eelc is the per bit energy dissipation in the radio electronics.

In this experiment, 100 sensor nodes randomly deployed in 100X100 m². The base station was positioned at (50, 175) m and the upper bound for number of clusters is set to 5% of nodes. The initial energy (d₀) of each sensor node was set to 2 J while the parameters utilized in the radio model are Eelc = 50 nJ/bit, efs = 10 pJ/bit/m² and emp = 0.0013 pJ/bit/m⁴. The microcontroller energy consumption for data aggregation (Eda) is assumed to be 5 nJ/bit/signal. The following assumptions are made throughout the experiment: error free communication channel, ideal MAC layer, and nodes are in range of each other and BS. Control packet size was set to 25 bytes, data packet size was set to 500 bytes, and 6 TDMA frames per each data gathering period was assumed.

Fig. 7 shows number of alive nodes per round for SimE and LEACH-C. The simulation was ended when number of dead nodes is greater than or equal to 90%. For LEACH-C, the first node died at round 512 and 90% of nodes died in round 1050. For SimE, the first node died at round 643 and 90% of nodes died in round 1151. Considering the network when 50% of the nodes died, SimE improves the lifetime by about 21% compared to LEACH-C. Energy consumption of the network over time is an important factor for measuring the efficiency of clustering the wireless sensor networks. In addition, the total energy consumption of network over time is shown in Fig. 8. The figure shows that SimE algorithm reduced the energy usage more than LEACH-C.

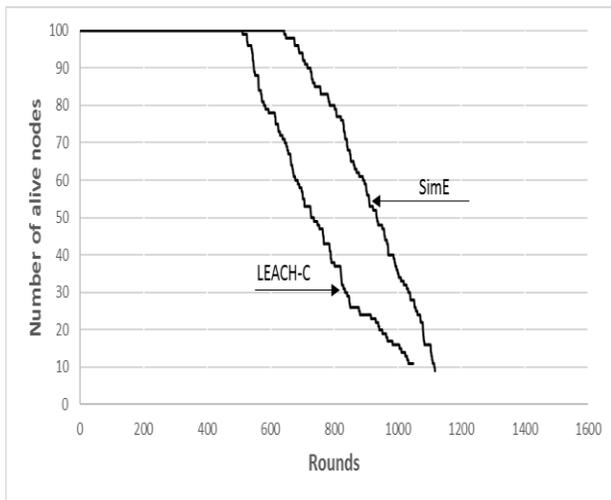


Fig. 7. Alive nodes vs. Simulation rounds.

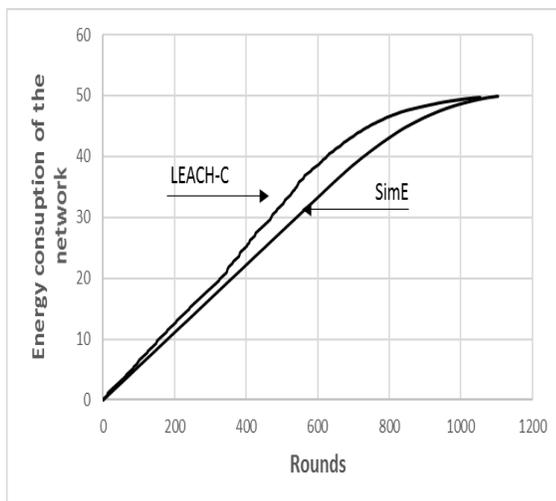


Fig. 8. Energy consumption of the network per round.

V. CONCLUSION

In this paper, cluster optimization in wireless sensor networks was presented using simulated evolution iterative algorithm. A goodness measure was proposed to evaluate the produced clusters. The proposed SimE approach and its goodness measure had the advantage of adaptively varying the clusters and their nodes when number of nodes in the network is decreased due to the complete loss of energy. This adaptivity is important for network lifetime as the nodes in a cluster might completely lose their energy in some round; causing unbalance in the produced clusters and re-clustering the whole network. Using adaptive SimE approach, the other clusters will remain unchanged and the whole network needs not to be re-clustered.

The results showed that SimE can produce a very high quality clusters for WSNs. In addition, the results showed that there is a relationship between the size of the deployment area, the number of clusters, and the reduction in the total distance. This suggests that number of clusters should be adaptive to number of alive nodes for better clustering in WSNs. Furthermore, the results demonstrated that the base station location is crucial for effectively clustering the WSN. Finally, the results depicted that the presented SimE approach increased the network lifetime by about 21% compared to LEACH-C protocol, which utilized SA algorithm as the base for selecting CHs.

REFERENCES

- [1] E. Otero, R. Haber, A. Peter, A. AlSayyari and I. Kostanic, "A Wireless Sensor Networks' Analytics System for Predicting Performance in On-demand Deployments," *IEEE Systems Journal*, vol. 9, no. 4, pp. 1344-1353, 2015.
- [2] A. AlSayyari, I. Kostanic and C. E. Otero, "An empirical path loss model for Wireless Sensor Network deployment in an artificial turf environment," in *Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control*, Miami, FL, USA, 2014.
- [3] I. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393-422, 2002.
- [4] A. Alsayyari, I. Kostanic and C. E. Otero, "An empirical path loss model for Wireless Sensor Network deployment in a concrete surface environment," in *2015 IEEE 16th Annual Wireless and Microwave Technology Conference (WAMICON)*, Cocoa Beach, FL, USA, 2015.
- [5] A. Alsayyari, I. Kostanic, C. E. Otero and A. Aldosary, "An empirical path loss model for wireless sensor network deployment in a dense tree environment," in *IEEE Sensors Applications Symposium (SAS)*, Glassboro, New Jersey, USA., 2017.
- [6] A. Alsayyari, I. Kostanic, C. Otero, M. Almeer and K. Rukieh, "An empirical path loss model for wireless sensor network deployment in a sand terrain environment," in *IEEE World Forum on Internet of Things (WF-IoT)*, Seoul, 2014.
- [7] P. Azad and V. Sharma, "Cluster Head Selection in Wireless Sensor Networks under Fuzzy Environment," *ISRN Sensor Networks*, vol. 2013, p. 8 pages, 2013.
- [8] W. B. Heinzelman, A. P. Chandrakasan and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on Wireless Communication*, vol. 1, no. 4, pp. 660-670, 2002.
- [9] V. Kumar, A. Kumar, Gaurav and M. Singh, "Improving network lifetime & reporting delay in wireless sensor networks using multiple mobile sinks," in *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016.
- [10] A. Jain and A. K. Goel, "Energy Efficient Algorithm for Wireless Sensor Network using Fuzzy C-Means Clustering," *International*

In conclusion, it is difficult to run a comprehensive comparison between the findings and performance results of this paper and other approaches proposed in the literature since the parameters utilized and the assumption made are different among the approaches. For instance, many proposals assumed a location for the BS inside or outside the network deployment area. As was illustrated in previous sections, the BS location is greatly influencing the simulations of the network. Also, other proposals assumed a fixed number of clusters (fixed CHs) in their simulations. Though, this assumption is completely avoided in our adaptive SimE approach and only an upper limit of number of clusters is used. However, some contributions assumed nearly the same assumptions made in LEACH/LEACH-C protocol. Looking at the distance reduction, however, the provided approach produced higher average distance reduction compared to [18] in most cases and CHs percentage is less. Comparing the presented SimE approach to other metaheuristics, SimE is about 8% better than PSO and SA algorithms presented in [23], which utilized mostly the same configuration and assumptions that were presented in this work.

- Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 4, 2018.
- [11] A. Ray and D. De, "Energy efficient clustering protocol based on K-means (EECPK-means)-midpoint algorithm for enhanced network lifetime in wireless sensor network," *IET Wireless Sensor Systems*, vol. 6, no. 181-191, p. 6, 2016.
- [12] D. Napoleon and P. G. Lakshmi, "An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points," in *Trendz in Information Sciences & Computing (TISC2010)*, Chennai, India, 2010.
- [13] F. Cao, J. Liang and G. Jiang, "An initialization method for the K-Means algorithm using neighborhood model," *Computers & Mathematics with Applications*, vol. 58, no. 3, pp. 474-483, 2009.
- [14] P. k. P. C. M. Agrawal, "Exact and Approximation Algorithms for Clustering," *Algorithmica*, vol. 33, no. 2, pp. 201-226, 2002.
- [15] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *the 33rd Annual Hawaii International Conference on System Sciences*, Maui, HI, USA, 2000.
- [16] R. Elhabyan and M. Yagoub, "PSO-HC: Particle swarm optimization protocol for hierarchical clustering in Wireless Sensor Networks," in *10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, Miami, FL, USA, 2014.
- [17] S. Kirkpatrick, C. D. Gelatt Jr. and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671-680, 1983.
- [18] S. Jin, M. Zhou and A. S. Wu, "Sensor network optimization using a genetic algorithm," in *7th World Multiconference*, 2003.
- [19] G. EkbataniFard, R. Monsefi, M.-R. Akbarzadeh-T and M. H. Yaghmaee, "A multi-objective genetic algorithm based approach for energy efficient QoS-routing in two-tiered Wireless Sensor Networks," in *IEEE 5th International Symposium on Wireless Pervasive Computing 2010*, Modena, Italy, 2010.
- [20] Yang, Zhengyu, et al. "Intermediate Data Caching Optimization for Multi-Stage and Parallel Big Data Frameworks." arXiv preprint arXiv:1804.10563, 2018.
- [21] Yang, Zhengyu, et al. "AutoTiering: automatic data placement manager in multi-tier all-flash datacenter." Performance Computing and Communications Conference (IPCCC), 2017 IEEE 36th International. IEEE, 2017.
- [22] N. M. Abdul Latiff, C. C. Tsimenidis and B. S. Sharif, "Energy-Aware Clustering for Wireless Sensor Networks Using Particle Swarm Optimization," in *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, Athens, Greece, 2007.
- [23] M. T. Mekonnen and K. N. Rao, "Cluster Optimization Based on Metaheuristic Algorithms in Wireless Sensor Networks," *Wireless Personal Communications*, vol. 97, no. 2, p. 1-15, 2017.
- [24] A. El Rhazi and S. Pierre, "A Tabu Search Algorithm for Cluster Building in Wireless Sensor Networks," *IEEE Transactions on Mobile Computing*, vol. 8, no. 4, 2009.
- [25] C. Boucetta, H. Idoudi and L. A. Saidane, "Hierarchical Cuckoo Search-based routing in Wireless Sensor Networks," in *IEEE Symposium on Computers and Communication (ISCC)*, Messina, Italy, 2016.
- [26] Yang, Zhengyu, et al. "A fresh perspective on total cost of ownership models for flash storage in datacenters." Cloud computing technology and science (CloudCom), 2016 IEEE International Conference on. IEEE, 2016.
- [27] R. Ostrovsky and Y. Rabani, "Polynomial time approximation schemes for geometric k-clustering," in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, Redondo Beach, CA, USA, 2000.
- [28] R. M. Kling and P. Banerjee, "ESp: Placement by simulated evolution," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 8, no. 3, pp. 245 - 256, 1989.
- [29] M. N. Halgamuge, M. Zukerman, K. Ramamohanarao and H. L. Vu, "An Estimation of Sensor Energy Consumption," *Progress in Electromagnetic Research*, vol. 12, pp. 259-295, 2009.

Investigating the Acceptance of Mobile Health Application User Interface Cultural-Based Design to Assist Arab Elderly Users

Ahmed ALsswey, Irfan Naufal Bin Umar, Brandford Bervell

Center for Instructional Technology and Multimedia, Universiti Sains Malaysia,
Pulau Penang, Malaysia

Abstract—Mobile health (m-health) applications are a way to provide solutions to the non-availability of physical health services in the Arab world. However, end users of m-health around the world have their cultural and personal differences that distinguish them from others. Studies suggest that culture is an essential component of the success of any product or technology usage. In view of this, the study investigated acceptance towards mobile health application User Interface (UI) designed for Arab elderly users based on their culture. The TAM model formed the theoretical basis upon which a quantitative design was adopted, with a questionnaire as data collection instrument from 134 participants. The findings showed that perceived Ease of Use (PEOU) and Attitude Towards Use (ATU) had a significant positive influence on Behavioural Intention (BI) to use mobile health application User Interface. Overall, the results indicated that Arab elderly users found the mobile health application UI as acceptable due to its cultural-based design. To improve designs of mobile applications UI targeting elderly users, it is vital to gain insight into cultural aspects that influence the usability of mHealth application UI as well as insights into their personal characteristics and experiences.

Keywords—TAM; elderly users; mobile health applications; user interface; culture

I. INTRODUCTION

Recently, the number of aged people over 65 years old has increased rapidly in the world and it is predicted to be about 1 billion by 2030 [1]. The significant increasing number of elders suggests the need to develop technologies for this group of users, in order to achieve their health demands [2], [3]. For example, the number of elderly generation aged 65 years and over in the United States is expected to rise from 40 million in 2010 to 72 million in 2030 [4]. This means that there should be more friendly and usable technologies developed to provide healthcare related support for this elderly population, as pressure on physical health facilities will be eminent.

Currently, it was reported in January 2018 that within the world, approximately 2.51 million times mobile health- apps (health and fitness) available in the Google Play Store were downloaded to mobile devices (<https://www.statista.com/statistics/699096/leadinghealth-and-fitness-google-play-canada-downloads/>). These health apps

involved numerous types of medical and health applications such as health and fitness apps for life management, medical education and patient-centered apps [5]. With the rapid increase in health apps, research on their use in the areas of healthcare evaluation, medicine and disease management have also increased meaningfully in the last years [6]. A large number of studies carried out on mobile apps concentrated on particular medical issues such as pain management [7], weight loss [8], [9], and diabetes [10], [11]. To analyze the content of applications and provide deep vision into what apps are presented and what practice based theories inform app design approach is of deep concern. In view of this, it is important to go behind content analysis to investigate users' perceptions towards m-health apps user interface (UI). Investigating users' experience with current m-health apps to give opportunity for researchers and apps developers to better design future m-health apps to be usable, effective and accepted by end users is becoming important. This is culturally possible when culture sophistication in design of such apps is to be addressed.

However, there is a lack of empirical studies that investigate and validate the effect of culture on UI design in Arab countries within the m-health context. Poor cultural aspects in UI design regularly means poor user interaction and hence reduced user acceptance and satisfaction. Therefore, addressing cultural differences of use in designing User Interfaces (UIs) could improve acceptance, usability and help users to interact in a better way with the interface. In addition, elderly users are not the main target of design technology and mobile applications [12]. This study contributes towards understanding Arab elderly users' attitudes towards using m-health application UI designed based on their cultural elements and the factors that influence users' intentions to use the designed mobile health application. In addition, recognizing and addressing these factors will support future design of mobile health applications UI and their implementation in the Arab world. The remainder of this paper is structured as follows. Section II describes the literature review of this study. Section III presents the research methodology. Section IV presents the results and analysis. Section V discusses the results of the study. Section VI presents conclusion and implications. Section VII describes the limitations of the study. Section VIII presents the future work.

II. LITERATURE REVIEW

A. Possible Barriers Faced by Elderly users in using Technology

One of the explanations for the seeming lack of m-health applications targeting Arab elders, is the lack of studies carried out to investigate the factors that affect elderly users' engagement in technology, to inform developers on the suitable m-health application appropriate for the Arab aged population [13]. On the other hand, there is the general idea in the world that the elderly are unwilling, unable and afraid to use technology involving mobile phones, internet and computers [14]-[16]. In addition, some elderly folks suffer from some effects of ageing such as hearing loss, psychomotor impairments, diminished vision as well as reduced attention, memory and learning abilities occurring at the beginning of ageing or resulting from some form of disease. These serve as barriers to interacting with technology. This may prevent them to effectively use technologies such as mobile applications, internet and navigating websites [17]. Around the world, the age group of 50 years and above suffer from some form of physical limitations and mental changes that can cause an interference with their engagement with technology [16]. In view of this, within the Arab context, there is the need to provide an m-health application coupled with an ideal environment that foster the engagement of elderly people aged 60 years and above with technology to improve upon their health behavior.

B. Elderly Users and Mobile Health Apps

According to [18], within The Netherlands, there has been an increasing interest among older people in using technology such as computers, smartphones and the internet. The search for health information and healthcare services is one of the most priorities for this group of users. One technology to deliver such health related services for elderly users is mobile technology, referred to as mobile health (m-Health). In addition, mobile health technologies have the ability to promote care to the elderly users through easy use and convenient healthcare. For instance, m-Health applications can remind patients to take medications, track daily pain levels, watch symptom levels and have access to behavioural health information [19, 20]. As the elderly lives longer, there is the need to improve their live through care and support by using technologies that adapt to ageing changes for better technology usable experience. M-health applications can provide such vital opportunities to learn and obtain health information to improve their lives [21].

C. Arab Elderly Users

The Arab world has witnessed rapid growth in the number of older adults. Based on the latest statistics in 2017, the older adults were 26.8 million in 2015 and expected to be 50 million by 2030. Presently, the percentage of elderly aged 60 years and above in the Arab world is expected to be 6.7%, with projections showing an increase to 9.5% by 2030 [22]. Therefore, the responsibility of the workforce is expected to change from the usual support for children, to the simultaneous support for both children and older persons [23]. This makes it

expedient to develop technologies targeted at Arab elderly users to primarily support their healthcare and reduce the envisaged dependency on the working class, while addressing the issue of non-availability of physical healthcare services. However, these mobile health technologies should be more convenient and acceptable for such people.

D. Arab Culture

Generally, end users around the world have their cultural and personal differences such as the customs, religion, habits and different languages that distinguish them from others. In addition, they interact using technology in different manners, depending on these differences [24]. In view of this, designing mobile applications' user interface to be more user friendly is a vital issue for the success of such applications, technologies and products. Nonetheless, the user interfaces in mobile phones face cultural differences, which mean that interfaces should be more convenient and acceptable to each cultural attribute [25].

The principles of Arab culture are special and unique, often unclear and secretive to Western cultures. Arab cultures are traditional, whereas the term Arab refers to language and cultural aspects, with many practices and norms revolving around Islamic beliefs and customs. So, everyday words, images, symbols or phrases that may be completely acceptable in western countries may be unacceptable for Arab audience [26]. Therefore, considering these differences in the UI design might be important for improved acceptance towards usability of UIs. This study designed a mobile health application UI based on Arab culture to enhance acceptance of using mHealth app UI.

E. Theoretical Framework and Development of Research Hypotheses

Several models have been suggested by researchers to investigate the individual acceptance behaviour on information technology and information systems, for examples of such models include Theory of Reasoned Action (TRA), [27], Theory of Planned Behaviour (TPB), [28]. Technology Acceptance Model (TAM, TAM2, and TAM3), [29], and Unified Theory of Acceptance and Use of Technology (UTAUT), [30]. Technology Acceptance Model (TAM) is one of the most popular theories used widely to investigate the individual acceptance behaviour of information systems [31]. TAM model as shown in Fig. 1, comprises variables such as behavioural intention to use, attitude towards use, perceived usefulness and perceived ease of use [32]. Davis (1989) defines perceived usefulness as the prospective user's subjective probability that using a specific application system will enhance his or her job or life performance, while perceived ease of use defines the degree to which an individual believes that using a particular system would be free of mental or physical effort. Additionally, behavioural intention is defined as "the strength of one's intention to perform a specified behaviour" [31]. Attitude on the other hand can be defined as "an individual's positive or negative feelings about performing a target behaviour" [33]. Previous reviews show that TAM may assist as a useful theoretical framework for the present study. Particularly, TAM provides its strength in predicting behavioural intention [34], [35].

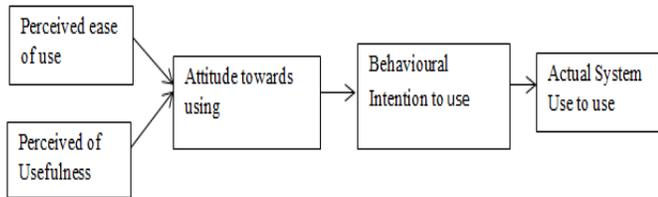


Fig. 1. Technology Acceptance Model (TAM) Davis (1989).

1) *Relationship between perceived ease of use (PEOU) and perceived usefulness (PU):* Perceived ease of use is expected to have a positive influence on perceived usefulness. The findings of the original TAM model have shown that perceived ease of use has a positive effect on perceived usefulness. Studies conducted by Gambari, and Sule [36], Liu [37] show that users do well in tasks when they do not need to exert much mental and physical effort. This study defines ease of use as the degree to which Arab elderly users' usage of mHealth app UI is perceived or seen as easy or effortless. Hence, this study hypothesizes that:

H1: Perceived ease of use will have a positive effect on user Perceived usefulness.

2) *Relationship between perceived usefulness (PU) and attitude:* Based on TAM studies, the first primary relationship is that perceived ease of use and perceived usefulness will have a possible impact on enhancing users' attitude towards usage. In addition, Juniwati [38] indicates that both perceived ease of use and perceived usefulness may affect attitude. Alsamydai [39] posited that, perceived usefulness was an important factor in using mobile banking services. It is supposed that perceived usefulness was affected by the level of users' trust [40]. Lai and Yang [41] claimed that users in a performance-oriented e-business environment are usually reinforced for good performance and benefits. Hence, this study defines usefulness as the degree to which Arab elderly users' usage of mHealth app UI is perceived as useful or beneficial. This notion will positively influence attitude toward mHealth app UI. Consequently, this study hypothesized that:

H2: Perceived usefulness will has positive effect on Attitude to use mHealth app UI.

3) *Relationship between perceived ease of use (PEOU) and attitude:* Attitude towards actual usage is determined by a belief of how easy the user thinks he can use the system. TAM posits that PEOU has a direct positive influence on attitude towards using a system. The present studies propose that perceived ease of use is a main attribute in defining the attitude of an individual towards system usage. This study defined attitude as the overall affection (whether positive or negative) of Arab elderly users towards using mHealth applications UI. Positive attitude is acquired when Arab elderly users find as easy when they use mHealth app UI. The reverse will be the case. This usually depends on the effort

needed to use the application and the complexity of the process. Therefore, this study hypothesizes that:

H3: Perceived ease of use will has positive effect on user attitude to use mHealth app UI.

4) *Relationship between perceived usefulness (PU) and behavioural intention (BI):* Davis [29] defined perceived usefulness (PU) as the users' belief that using a particular system will increase his or her job performance. Based on this definition, Phua, Wong, and Abu [42] found PU as a main factor of usage behaviour and intention. Subramanian [43] found that PU (not PEOU), had a direct impact on usage behaviour. This study, define PU as the degree to which Arab elderly users believe that the use of mHealth app UI will improve their health related needs or livelihood. Consequently, this study hypothesizes that:

H4: Perceived usefulness will has positive effect on user behavioural intention to use mHealth app UI.

5) *Relationship between perceived ease of use (PEOU) and behavioural intention (BI):* Perceived ease of use (PEOU) has an indirect impact on behavioural intention to use technology through increased perceived usefulness [44]. The easier a system is to interact with; the greater should be the users' sense of efficacy [45] and personal control [46], concerning his or her ability to move out the sequences of behaviour required to run the system. This study defines PEOU as, flexibility, simplicity and compatibility towards the overall usage of the mobile health application as experienced by Arab elderly users. Subsequently, the variables attached to perceived ease of use (flexibility, simplicity and compatibility) were all positively related to behavioural intention in the literature. Hence, this study hypothesizes that:

H5: Perceived ease of use will have a positive effect on behavioural intention to use mHealth app UI.

6) *Relationship between attitude and behavioural intention (BI):* Attitude (ATU) is a vital construct in the domain of information technology. In terms of mobile applications, most users today prefer to use mobile apps for various purposes because of their portability characteristics. Based on their previous experiences, they develop an attitude towards using them, ranging from good to poor. Previous empirical studies have shown the existence of such generalized attitude and its effects on the assessment of new systems in similar situations [47]-[49]. In this study, attitude is envisaged to have effects on intention towards using mHealth app UI. The variable is defined in this study as the degree to which Arab elderly users have favourable or unfavourable affection towards using mHealth app UI. Numerous studies such as that of Shittu, Gambari, and Sule [36] and Wang and Liu [50] have stated that the attitude construct affect behavioural intention. This study posits that how favourable or unfavourable Arab elderly users perceive the use of mHealth applications UI to be, will eventually influence their intention behaviour to use the application. Against this backdrop, they study hypothesizes that:

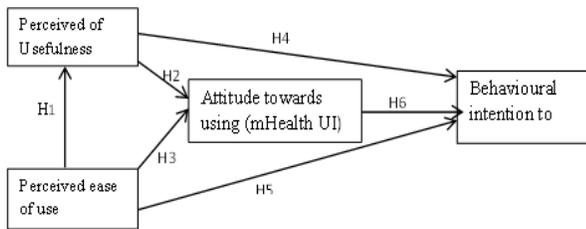


Fig. 2. Conceptual framework.

H6. Attitude towards use will have a positive effect on behavioural intention to use mHealth app UI. The conceptual framework based on this discussion is presented in Fig. 2.

F. Mobile Health Application based on Arab Culture

A mobile health application was specifically developed for Arab elderly users for this study. The application manages their health related information and needs such as dosage, time, type of medicine and instructions about medication. In addition, this application provides general health related information on the causes, symptoms and preventive strategies of diseases specific or common to the Arab world. Fig. 3 depicts a graphical image of the interfaces for this mobile health application.

Arabic language used in designing this app UI was to ensure that the functions are clear and understandable for Arab elderly users from different backgrounds to avoid any confusion that can occur due to different meaning of some local terms and words. Colours of Arab culture and Islam (green, blue and black) are used in designing app UI to enhance acceptance of Arab users, since Arab people are proud of their culture, customs, and religion. Green reflects the Islamic symbol, while black is linked to a specific period of time in the Islamic era. The Blue colour is related to the sea and sky. Red and yellow colours also used in designing the application UI were to attract the users. Since these colours are used frequently in the Arab world to get the attention of users and attract them to try its usage and buy different things such as clothes, foods, cars etc.

The app UI used font size of 12pt and font type “الرقعة” to display the information while font size 14pt was used for heading. This is because it is one of the easiest Arabic fonts and the most widespread among people in their daily writings. Additionally, its simplicity and distance from the complexity and ease of reading and writing are very much appreciate in the Arab world [51]. Information architecture of app UI was designed simply based on previous studies conducted on Arab culture such as by Hall [52] and Hofstede [53]. They showed that Arab culture is Uncertainty Avoidance culture, which means that Arab people do not like risk, and prefer simple use and avoid any complex design of systems.

Arab culture has strong Individualism and Collectivism dimension [52], [53]. This means that Arab users have high concern for the group and exchange for devotion. Reputation, dignity, shame, honour, and pessimism occupy a higher consideration. Therefore, common Arab icons and symbols are used in the app UI design. In addition, Arab writing and

reading are from right to left. Therefore, the UI layout was designed from right to left. Finally, labels and messages were used in the app UI design to inform users about the various stages of usage and completed tasks. This was targeted at providing easy assistance and information on progress made with the system’s use.



Fig. 3. Some Screen Shoot of the Designed mHealth App.

III. RESEARCH METHODOLOGY

A. Research Method

1) *Questionnaire*: The questionnaire for this study was adopted from Davis (1989) and modified to suit mobile health app UI context which have also been validated in previous studies such as Ntaliani, Costopoulou [54], Alharbi and Drew [32], to guarantee clarity and avoid grammatical and language errors, the questionnaire was first examined by English language experts and then translated into Arabic language by Arabic language experts. This provided clarity and accuracy of understanding to the respondents. This is because the respondents were native Arabic speakers.

2) *Instrument*: The questionnaire consisted of five parts. Part one focused on demographic characteristics of participants. Parts two, three, four and five elicited responses on factors related to TAM model constructs. Specifically, parts two and three measured participants’ Usefulness, Ease of Use toward using m-health application UI respectively, while part four measured the factors that impact users’ Attitude and part five on Behavioural Intention to use mobile health app UI. All TAM factors were measured at five levels of Likert-type scale. Participants were asked to choose from 5 points Likert-type scale with 1= SD (Strongly Disagree), 2= D (Disagree), 3= N (Not Sure), 4= A (Agree) and 5= SA (Strongly Agree).

3) *Sample and data collection*: In this study, the target population was Arab elderly users (male and female) from different backgrounds aged 60 years and above and have at least one year experience in using mobile applications. This study used purposive sample technique to obtain the sample size for this study. 150 questionnaires were distributed but a total of 137 questionnaires were received from the targeted respondents, which constituted 91.3% response rate from the survey. Among the 137 sets of questionnaires returned, there were incomplete responses from four respondents, leaving 134 questionnaires for further data analysis use.

IV. DATA ANALYSIS

A. Demographics

Demography of participants for the study was analysed by simple descriptive statistics. Results from this analysis are depicted in Table I.

TABLE I. DEMOGRAPHIC INFORMATION STATISTICS FOR PARTICIPANTS

	Information	Number of participants	Percentage of sample
Age	60-64	118	88.1%
	65-69	10	7.5%
	70-74	4	3.0%
	75-79	2	1.5%
	≥80	0	0%
Participants' Gender	Male	113	84.3%
	female	21	15.7%
Participants' Level of Education	School level		
	Diploma degree	58	43.3%
	Bachelor degree	44	32.8%
	Master degree	15	11.2%
	PhD degree	8	6.0%
Participants' experience level of mobile apps use	1-3	9	6.7%
	4-6	59	44.0%
	7-9	46	34.3%
	≥10	14	10.4%

As shown in Table I, majority of respondents were between the age range of 60-64 years (n: 118; 88.1%). However few of them (n: 2; 1.5%) were between 75-79 years old. None of the participant was 80 years or above. Regarding the gender, out of the 134 respondents, 113 of them (84.3%) are male and 21 (15.7%) are female. With regards to the level of education, most of the respondents had school level certificate (n: 58; 43.3%). A few of them possessed masters (n: 8; 6.0%) and Ph.D. (n: 9; 6, 7%). In terms of experience in mobile application usage, majority of the users.

B. Instrument Reliability

Reliability is the consistency of measurement or stability of measurement over a variety of conditions in which basically the same results should be obtained [55]. The reliability analysis was based on the Cronbach's Alpha tests on the items used for each factor. Table II shows the reliability of each factor of this study.

TABLE II. RESULTS OF RELIABILITY ANALYSIS

Factor	Items	Cronbach's alpha
Perceived Usefulness (PU)	6	.878
Ease of use (EOU)	7	.796
Attitude (AT)	4	.725
Behavioural Intention (BI)	3	.703
Overall reliability of Acceptance factors	20	.897

Reliability values can be classified into four ranges: values of up to 0.50 indicate low reliability; values from 0.50 to 0.70 indicate moderate reliability; and values from 0.70 to 0.90 indicate high reliability respectively [56]. The findings shown in Table II indicate that all reliability values were higher than 0.70 which is considered as high. Moreover, when all items were entered at the same time, the overall reliability for this instrument was 0.897, exceeding the acceptable threshold.

C. Acceptance Correlation Analysis

The correlation coefficients were examined to determine the relationship between Perceived ease of use (PEOU), Perceived Usefulness (PU), Attitude (AT) and Behavioral Intention (BI) toward mobile health application UI usage. Pearson's correlation analysis was employed, since it is the most common measure of correlation which shows the degree of linear relationship between variables and it is useful for scale and interval variable relationships [57]. Table III shows the correlations between the PEOU, PU, AT, and BI.

Table III shows that the correlations between acceptance factors PEOU, PU, AT and BI are significant and positive. Based on Pearson correlation 1 to 0.3 represents small; 0.3 to 0.5 medium; and 0.5 to 1.0 large [58]. Table III shows there is a moderate significant positive correlation between perceived ease of use and perceived usefulness (r = 0.379**, p<0.01). Additionally, the results indicated that there is a moderate significant positive correlation between perceived usefulness and attitudes towards using the mobile health application UI (r = 0.345**, p<0.01). Furthermore, the findings show that there is a large significant positive correlation between perceived ease of use and attitudes towards using the mobile health app UI (r = 0.760**, p <0.01). Similarly, the relationship between attitude towards using the mobile health app and Behavioral Intention was significantly positive and large (r = 0.641**, p<0.01).

TABLE III. PEARSON CORRELATION BETWEEN VARIABLES

Factors	Behavioural Intention (BI)	Attitude (AT)	Ease of use (EOU)	Perceived Usefulness (PU)
Behavioural Intention (BI)	1	.641** .000	.622** .000	.322** .001
Attitude (AT)	.641** .000	1	.760** .000	.345** .001
Ease of use (PEOU)	.622** .000	.760** .000	1	.379** .000
Perceived Usefulness (PU)	.322** .001	.345** .001	.379** .000	1

D. Factors Affecting Arab Elderly Users' Behavioural Intention of using Mobile Health Application UI

To find the factors affecting Arab elderly users' acceptance of using mobile health application UI, stepwise multiple regression was conducted. Before continuing for the regression analysis, suitability of the regression analysis was assessed to ensure that there was no violation of the assumptions of outliers, normality, multi-collinearity, homoscedasticity, linearity, and independence of residual. Based on Kock [59], in assessing multi-collinearity among variables, the best approach is to analyze the variance inflation factor (VIF) values. Accordingly, VIF values below 3.0 indicate the absence of multi-collinearity. Based on the result for this study, all VIF values for the relationships between dependent and independent variables ranged from 1.000 to 2.258 which are below the 3.0 threshold [61]. The VIF values for dependent variables are shown in Table IV.

TABLE IV. MULTI-COLLINEARITY STATISTICS VIF

Independent Variables	Dependent Variable		
	Behavioural	Attitude	Usefulness
Ease of use	1.857	1.065	1.000
Usefulness	1.312	1.065	
Attitude	2.258		

E. Multiple Regressions between Perceived Ease of Use, Usefulness, Attitude and Behavioural Intention

A multiple regression was finally applied to test for predictive significance between dependent and independent variables. The results of the stepwise regression are presented in Table V.

TABLE V. ANOVA RESULTS BETWEEN PERCEIVED EASE OF USE, USEFULNESS, ATTITUDE AND BEHAVIOURAL INTENTION

No.	Model	Sum of Squares	df	Mean Square	F
1.	Regression	6.331	1	6.331	55.134
	Residual	9.071	79	.115	
	Total	15.402	80		
2.	Regression	6.992	2	3.496	32.424
	Residual	8.410	78	.108	
	Total	15.402	80		

Dependent Variable: Behavioral Intention
 Predictors: (Constant), Attitude
 Predictors: (Constant), Attitude, Ease of use

Table V shows ANOVA, a test of significance of model. The results show that two predictor variables (ease of use and attitude) are a statistically significant predictors of behavioural intention in using mobile health application UI. Table VI shows the values of regression coefficients of the two regression models constructed by stepwise regression method. Two independent variables emerged as positive predictors of behavioural intention.

TABLE VI. VALUES OF MULTIPLE LINEAR REGRESSION BETWEEN PERCEIVED USEFULNESS, PERCEIVED EASE OF USE, ATTITUDE AND BEHAVIOURAL INTENTION

Model	Unstandardized Coefficients		Standardized Coefficients	t	sig
	B	Std. Error	Beta		
Behavioural Intention	1.394	.964		3.777	.000
(Constant)	.629	.391	.641	7.425	.000
Attitude	.964	.397		2.425	.018
Ease of use	.391	.126	.399	3.094	.003
(Constant)	.344	.139	.319	2.476	.015
Usefulness	.255	.084	.322	3.021	.003
(Constant)					

a. Dependent Variable: Behavioural Intention to use

The Behavioural Intention of Arab elderly users was primarily determined in a positive manner by Ease of Use ($\beta = 0.319, p < 0.05$), Attitude ($\beta = 0.399, p < 0.05$), and Usefulness ($\beta = 0.322, p < 0.05$). Apparently, Attitude was the strongest predictor in affecting the Behavioural Intention in using mobile health application user interface. The Attitude of Arab elderly users was in turn influenced by ease of use ($\beta = 0.319, p < 0.05$) and strongly determined by Usefulness ($\beta = 0.322, p < 0.05$). However Arab elderly users' perception of the usefulness of the mobile health app was based on easiness of use of the application ($\beta = 0.322, p < 0.05$).

V. DISCUSSION

The findings of this study showed that Arab elderly users have positive intentions to use the designed mHealth app UI. TAM constructs (attitude, usefulness, and ease of use) had high prediction levels in explaining behavioural intention to use mobile health app UI. Ultimately, the construct attitude was the strongest predictor of the mHealth UI app uptake intention. This means that the sample population had a positive attitude toward using mobile health app UI which was designed based on their culture. This induced positive attitudes towards their use of the application in the future. Consequently, the majority of Arab elderly users believed that using mobile health app UI was also positive and a good idea. This finding supports studies such as that of Alotaibi, Houghton [60] and Fong and Wong [61] who found that the mainstream of participants wanted to use mobile application and services based on their positive attitude toward the application.

Additionally, ease of use had a strong significant positive relationship towards attitude to use the mobile health app. This indicates that the mobile health app UI design was clear, simple and easy to use. This was because the consistency between UI elements designed based on Arabic aspects such as language, layout, colors, and buttons promoted ease of use. These findings are supported by several studies that found perceived ease of use to have a close correlation to the construct attitude in TAM [42], [62]. In addition, the result of this study is supported by other studies conducted on the role

of culture in technology design such as that of Almakky, Sahandi [63], Ishak, Jaafar [64] and Kalliny, Saran [26]. These previous researchers also addressed the fact that users feel it is easy to interact with technologies and user interfaces which directly reflect and relate to their own culture. Furthermore, perceived ease of use had a strong significant relationship with behavioural intention. Arab elderly users believed that the design of mobile health app UI was easy to use and effective to use. This usage would be expected to strengthen their perception of the easiness of mHealth app UI. The flexibility in obtaining information, operating the application, and learning and managing medical information, provides a sense of relief to Arab elders on less effort exerted towards usage. Consequently, Agarwal and Karahanna [65] and Venkatesh [66], reiterate that successful deployment of mobile applications are based on how developers and designers ensure easiness towards actual use of apps Redzuan, Razali [67] and [68] also provided evidence that perceived ease of use has a significant effect on behavioural intention to use systems.

The relationship between ease of use and usefulness was also positive and significant within this study. The effect of perceived ease of use might contribute to clarifying a significant impact towards perceived usefulness. Usefulness of an application is tied to actual use. Where users are able to apply a particular technology to perform associated tasks, they relate to the performance of the technology in meeting task requirements. The ability of Arab elders in utilizing the mobile health app UI for health related purposes provided them with a sense of usefulness of the app. This is because, they derived benefits from the app in terms of acquiring useful health related information, diagnosis, prescriptions and awareness of ageing ailments' management and prevention. The acquisition of these benefits from the mobile app usage was dependent on their ability to interact with the app when demanded. Davis [28], posited that ease of use is an originator of usefulness. This result is also consistent with previous studies from Lallmahamood [69], Leishman [70], as well as Shim and Viswanathan [71].

Finally, the construct usefulness significantly influenced attitude towards usage of the mobile health app. Arab elderly users were interested in using the functionality of the mobile health app UI, which in turn increased their health related independence and helped them in their daily activities, ultimately improving their healthy life. The importance attached to the provisions of the mobile health app UI, made the users acquire a favourable affection towards the app. Their feelings were positive towards using the app because it benefited them in terms of provision of their health needs. In addition, usefulness had a significant effect on behavioural intention to use the mobile health app UI. As users became familiar with the use of the mobile health app, they found it helpful in their lives. How helpful the app was to them, culminated into positive intentions towards continual usage of the app. Arab elderly users' behavioural intentions to use the app was positive and highly reflected in their final usage behaviour. This result confirms the original TAM relationship between perceived usefulness and intention to adopt new technology. The result is also in line with several studies such

as by [29, 30] who stated that usefulness was a strong factor for measuring the behavioural intention to use new systems.

VI. CONCLUSION AND IMPLICATIONS

Mobile health apps could be a vital part for developing elderly people lives and decrease the cost of healthcare. This study attempted to enlarge the lack of research on the Arab elderly users' perspectives of mobile health app UI when culture is integrated into the design of mHealth application UI. Theoretically, this study evaluated the technology acceptance model (TAM) in the mobile health app UI context by investigating relationships between behavioural intention to use (BIU) and three independent factors perceived ease of use (PEOU), perceived usefulness (PU), and attitude towards use (ATU). The study identified the significant factors as being PU, PEUO and ATU. The results showed that most participants had a positive attitude towards using mobile health application UI designed based on Arab culture. However, this study provides support for TAM, and approved the fact that attitude is most powerful factor in explaining the behavioural intention toward usage.

VII. LIMITATIONS

Nonetheless, there are some limitations that should be taken into account. First, the study's conceptual framework was constructed based on TAM only. This study can be further improved by other theoretical frameworks like Unified Theory of Acceptance and Use of Technology (UTAUT) to add to the analytical power of acceptance. Second, the present study investigated only the determinants of behavioural intention without exploiting that of actual behaviour.

VIII. FUTURE WORK

In the future, the researcher plans to continue exploring the relationship between certain design elements of UI, cultural aspects and mobile applications. Of particular interests are the following cultural aspects: Age, gender, religion and education level. This further exploration will be achieved by conducting future studies using mobile applications and theoretical frameworks like Unified Theory of Acceptance and Use of Technology (UTAUT).

REFERENCES

- [1] Mamolo, M. and S. Scherbov, Population projections for forty-four European countries: The ongoing population ageing. 2009: Vienna Inst. of Demography.
- [2] Roupa, Z., Nikas, M., Gerasimou, E., Zafeiri, V., Giasyrani, L., Kazitori, E., & Sotiropoulou, P. (2010). The use of technology by the elderly. *Health Science Journal*, 4(2)..
- [3] Boustani, S., Designing touch-based interfaces for the elderly. University of Sydney, Sydney, 2010.
- [4] Vincent, G. and V. Velkoff, US Census Bureau. The next four decades: the older population in the United States: 2010 to 2050: population estimates and projections. May 2010. 2013.
- [5] Shih, S.P., S. Yu, and H.C. Tseng, The Study of Consumers' Buying Behavior and Consumer Satisfaction in Beverages Industry in Tainan, Taiwan. *Journal of Economics, Business and Management*, 2015. 3(3): p. 391-394.
- [6] Boulos, M.N.K., et al., How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX. *Biomedical engineering online*, 2011. 10(1): p. 24.

- [7] Stinson, J.N., et al., Development and testing of a multidimensional iPhone pain assessment application for adolescents with cancer. *Journal of medical Internet research*, 2013. 15(3).
- [8] Carter, M.C., et al., Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. *Journal of medical Internet research*, 2013. 15(4).
- [9] Turner-McGrievy, G.M., et al., Comparison of traditional versus mobile app self-monitoring of physical activity and dietary intake among overweight adults participating in an mHealth weight loss program. *Journal of the American Medical Informatics Association*, 2013. 20(3): p. 513-518.
- [10] Khirwan, M., et al., Diabetes self-management smartphone application for adults with type 1 diabetes: randomized controlled trial. *Journal of medical Internet research*, 2013. 15(11).
- [11] Cafazzo, J.A., et al., Design of an mHealth app for the self-management of adolescent type 1 diabetes: a pilot study. *Journal of medical Internet research*, 2012. 14(3).
- [12] Khaddam, I., & Vanderdonck, J., Towards a Culture-Adaptable User-Interface Architecture. *Romanian Journal of Human-Computer Interaction*, 2014. 7(2), 161.
- [13] Kuerbis, A., et al., Older adults and mobile technology: Factors that enhance and inhibit utilization in the context of behavioral health. 2017.
- [14] Foreman, K.F., et al., Impact of a text messaging pilot program on patient medication adherence. *Clinical therapeutics*, 2012. 34(5): p. 1084-1091.
- [15] Olson, K.E., et al., Diffusion of technology: frequency of use for younger and older adults. *Ageing international*, 2011. 36(1): p. 123-145.
- [16] Hanson, V.L., Technology skill and age: what will be the same 20 years from now? *Universal Access in the Information Society*, 2011. 10(4): p. 443.
- [17] Azir Rezha, N., Maksom, Z., & Naim, C. P. Tackling design issues on elderly smartphone interface design using activity centered design approach. *ARPN J. Eng. Appl*, 2014. Sci, 9, 1190-1196.
- [18] Alpay, L., et al., Current challenge in consumer health informatics: Bridging the gap between access to information and information understanding. *Biomedical informatics insights*, 2009. 2: p. BII. S2223.
- [19] Mercer, K., et al., Using a collaborative research approach to develop an interdisciplinary research agenda for the study of mobile health interventions for older adults. *JMIR mHealth and uHealth*, 2015. 3(1).
- [20] Levine, M. and M. Reid, D45: Primary care providers' perspectives on telemedicine in the pharmacologic management of older adults with chronic pain (cp). *Journal of the American Geriatrics Society*, 2012. 60: p. S202.
- [21] Mitzner, T.L., et al., Older adults talk technology: Technology usage and attitudes. *Computers in human behavior*, 2010. 26(6): p. 1710-1721.
- [22] Sibai, A.M., et al., Ageing and health in the Arab region: Challenges, opportunities and the way forward. *Population Horizons*, 2017.
- [23] Saxena, P.C., Ageing and age-structural transition in the Arab countries: Regional variations, socioeconomic consequences and social security. *Genus*, 2008: p. 37-74.
- [24] Khaddam, I. and J. Vanderdonck, Towards a Culture-Adaptable User-Interface Architecture. *Romanian Journal of Human-Computer Interaction*, 2014. 7(2): p. 161.
- [25] Van Biljon, J. and P. Kotzé, Cultural factors in a mobile phone adoption and usage model. *J. UCS*, 2008. 14(16): p. 2650-2679.
- [26] Kalliny, M., et al., Cultural differences and similarities in television commercials in the Arab world and the United States. *Journal of global marketing*, 2011. 24(1): p. 41-57.
- [27] Fishbein, M. and I. Ajzen, *Belief, attitude, intention and behavior: An introduction to theory and research*. 1975.
- [28] Fishbein, M. and I. Ajzen, *Belief, attitude, intention, and behavior: An introduction to theory and research*. 1977.
- [29] Davis, F.D., Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 1989: p. 319-340.
- [30] Venkatesh, V., et al., User acceptance of information technology: Toward a unified view. *MIS quarterly*, 2003: p. 425-478.
- [31] Al-Hujran, O. and M. Al-Dalameh. The role of national culture on citizen adoption of e-government web sites. in *ECEG2011-Proceedings of the 11th European Conference on EGovernment: ECEG2011*. 2011. Academic Conferences Limited.
- [32] Alharbi, S. and S. Drew, Using the technology acceptance model in understanding academics' behavioural intention to use learning management systems. *International Journal of Advanced Computer Science and Applications*, 2014. 5(1): p. 143-155.
- [33] Al-Adwan, A., A. Al-Adwan, and J. Smedley, Exploring students acceptance of e-learning using Technology Acceptance Model in Jordanian universities. *International Journal of Education and Development using Information and Communication Technology*, 2013. 9(2): p. 4.
- [34] Wingo, N. P., Ivankova, N. V., & Moss, J. A. Faculty Perceptions about Teaching Online: Exploring the Literature Using the Technology Acceptance Model as an Organizing Framework. *Online Learning*, 2017. 21(1), 15-35.
- [35] Lai, P. C. The literature review of technology adoption models and theories for the novelty technology. *JISTEM-Journal of Information Systems and Technology Management*, 2017. 14(1), 21-38.
- [36] Shittu, A. T., Gambari, A. I., & Sule, A. O. Students' attitude and behavioural intention on adoption of Internet for learning among Al-Hikmah University Students in Nigeria: A test of technology acceptance model. *Malaysian Journal of Distance Education*, 2013.15(2), 89-107. Ocean University, 72
- [37] Liu, Z. Y. An Analysis of Technology Acceptance Model-Exploring user acceptance and intension of taxi-hailing app in Shanghai, 2015. p. 206-215.
- [38] Juniwati, J. Influence of perceived usefulness, ease of use, risk on attitude and intention to shop online. *European Journal of Business and Management*, 2014. 6, 218-229
- [39] Alsamydai, M. J. Adaptation of the technology acceptance model (TAM) to the use of mobile banking services. *International Review of Management and Business Research*, 2017. 3(4), 2039.
- [40] Guritno, S., & Siringoringo, H. Perceived usefulness, ease of use, and attitude towards online shopping usefulness towards online airlines ticket purchase. *Procedia-Social and Behavioral Sciences*, 2013. 81, 212-216.
- [41] Lai, J.-Y. and C.-C. Yang, Effects of employees' perceived dependability on success of enterprise applications in e-business. *Industrial Marketing Management*, 2009. 38(3): p. 263-274.
- [42] Phua, P. L., Wong, S. L., & Abu, R. Factors influencing the behavioural intention to use the internet as a teaching-learning tool in home economics. *Procedia-Social and Behavioral Sciences*, 2012.59, 180-187.
- [43] Rathore, S., & Panwar, A. Factors Influencing Behavioural Intention to use Smart Phones. *Global Vistas*, 2015. 19-28.
- [44] Sek, Y.-W., et al., Prediction of user acceptance and adoption of smart phone for learning with technology acceptance model. *Journal of Applied Sciences(Faisalabad)*, 2010. 10(20): p. 2395-2402.
- [45] Redzuan, N. I. N., Razali, N. A., Muslim, N. A., & Hanafi, W. N. W. Studying Perceived Usefulness and Perceived Ease of Use of Electronic Human Resource Management (e-HRM) with Behavior Intention. *International Journal of Business*, 2016.1(2).
- [46] Lepper, M.R., Microcomputers in education: Motivational and social issues. *American Psychologist*, 1985. 40(1): p. 1.
- [47] Moon, J.-W. and Y.-G. Kim, Extending the TAM for a World-Wide-Web context. *Information & management*, 2001. 38(4): p. 217-230.
- [48] O'Cass, A. and T. Fenech, Web retailing adoption: exploring the nature of internet users Web retailing behaviour. *Journal of Retailing and Consumer services*, 2003. 10(2): p. 81-94.
- [49] Wang, Y.-F. Modeling predictors of restaurant employees' green behavior: Comparison of six attitude-behavior models. *International Journal of Hospitality Management*, 2016.58, 66-81.
- [50] Wang, W.-H. and Y.-J. Liu, Attitude, behavioral intention and usage: An empirical study of Taiwan Railway's internet ticketing system. Taiwan: National Taiwan Ocean University, 2009: p. 72.
- [51] Azmi, A. and A. Alsaiani, Arabic typography: a survey. *International Journal of Electrical & Computer Sciences*, 2010. 9(10): p. 1.

- [52] Hall, The theory of groups. Vol. 288. 1976: American Mathematical Soc.
- [53] Hofstede, Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. 2001: Sage.
- [54] Ntaliani, M., C. Costopoulou, and S. Karetos, Mobile government: A challenge for agriculture. *Government Information Quarterly*, 2008. 25(4): p. 699-716.
- [55] Drost, E.A., Validity and reliability in social science research. *Education Research and perspectives*, 2011. 38(1): p. 105.
- [56] Cho, E., & Kim, S. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 2015.18(2), 207-230.
- [57] Hauke, J., & Kossowski, T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 2011. 30(2), 87-93.
- [58] Field, A., *Discovering statistics using IBM SPSS statistics*. 2013: sage.
- [59] Kock, N., Non-normality propagation among latent variables and indicators in PLS-SEM simulations. *Journal of Modern Applied Statistical Methods*, 2016. 15(1): p. 16.
- [60] Alotaibi, R., L. Houghton, and K. Sandhu, Factors Influencing Users' Intentions to Use Mobile Government Applications in Saudi Arabia: TAM Applicability. *international journal of advanced computer science and applications*, 2017. 8(7): p. 200-211.
- [61] Fong, K.K.-K. and S.K.S. Wong, Factors influencing the behavior intention of mobile commerce service users: An exploratory study in Hong Kong. *International Journal of Business and Management*, 2015. 10(7): p. 39.
- [62] Burton-Jones, A. and G.S. Hubona, Individual differences and usage behavior: revisiting a technology acceptance model assumption. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 2005. 36(2): p. 58-77.
- [63] Almakky, H., R. Sahandi, and J. Taylor, The Effect of Culture on User Interface Design of Social Media-A Case Study on Preferences of Saudi Arabians on the Arabic User Interface of Facebook. *World Academy of Science, Engineering and Technology International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 2015. 9(1): p. 107-111.
- [64] Ishak, Z., A. Jaafar, and A. Ahmad, Interface design for cultural differences. *Procedia-Social and Behavioral Sciences*, 2012. 65: p. 793-801.
- [65] Agarwal, R. and E. Karahanna, Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, 2000: p. 665-694.
- [66] Venkatesh, V., Creation of favorable user perceptions: exploring the role of intrinsic motivation. *MIS quarterly*, 1999: p. 239-260.
- [67] Redzuan, N.I.N., et al., Studying Perceived Usefulness and Perceived Ease of Use of Electronic Human Resource Management (e-HRM) with Behavior Intention. *International Journal of Business*, 2016. 1(2).
- [68] Revythi, A. and N. Tselios, Extension of Technology Acceptance Model by using System Usability Scale to assess behavioral intention to use e-learning. *arXiv preprint arXiv:1704.06127*, 2017.
- [69] Lallmahamood, M., An Examination of Individual's Perceived Security and Privacy of the Internet in Malaysia and the Influence of this on their Intention to Use E-commerce: Using an Extension of the Technology Acceptance Model. *Journal of Internet Banking and Commerce*, 2007. 12(3): p. 1.
- [70] Leishman, P., Understanding the Unbanked Customer and Sizing the Mobile Money Opportunity. *GSMA Mobile Money for the Unbanked Annual Report 2009, 2010*.
- [71] Shim, S.J. and V. Viswanathan, User assessment of personal digital assistants used in pharmaceutical detailing: system features, usefulness and ease of use. *Journal of Computer Information Systems*, 2007. 48(1): p. 14-21.

Acoustic Classification using Deep Learning

Muhammad Ahsan Aslam, Muhammad Umer Sarwar, Muhammad Kashif Hanif, Ramzan Talib, Usama Khalid
Department of Computer Science
Government College University
Faisalabad, Pakistan

Abstract—Acoustic complements is an important methodology to perceive the sounds from environment. Significantly machines in different conditions can have the hearings capability like smartphones, different software or security systems. This kind of work can be implemented through conventional or deep learning machine models that contain revolutionized speech identification to understand general environment sounds. This work focuses on the acoustic classification and improves the performance of deep neural networks by using hybrid feature extraction methods. This study improves the efficiency of classification to extract features and make prediction of cost graph. We have adopted the hybrid feature extraction scheme consisting of DNN and CNN. The results have 12% improvement from the previous results by using mix feature extraction scheme.

Keywords—Acoustics; deep learning; machine learning; neural networks; audio sounds

I. INTRODUCTION

The advances in the automatic recognition of voice were consolidated in industrial systems [1]. Researchers have more interest to make advancement in identification quality. It is challenging task to identify acoustics in remote situations against the noisy background [2]. In other places, the advances in retrieval of music information have provided us with systems that can transcribe the notes and chords in the music or check the name of the track and the artist from a fragment of low quality sounds. The main focus of researchers is on classification of speech and music which can be heard mostly in a typical indoor outdoor environment [3]. Sound is few times a purposeful completion of different methods such as video, which transports information that would not be present like speech, processed data and songs of birds. The voice may also be easier to collect on a cellular phone. The information collected from a pragmatics audio inquiry can be purposeful for more working such as machine exploration, alert messages for user or analysis and prediction of event arrangements [4].

In the past years, a number of sound control techniques have been proposed. Deep learning is possibly the most recently used. The term deep learning employs a high level representation of low level data by stacking multiple levels using nonlinear module. There are several variants of deep learning architectures. The convolutional neural network is a profound learning method traditionally used for image distinction because of its good performance in learning distinctive local characteristics. The first Detection and Classification of Acoustic Scenes and Events (DCASE) challenge organized by the IEEE Audio and Acoustic Signal

Processing (AASP) in 2013 and then the DCASE 2016 challenge with an extend Acoustic scene classification dataset [5].

A. Acoustic Scene Classification

Acoustic is a term used in different fields. Acoustics is challenged in many terms, as in acoustic physics means the knowledge in the field of mechanical waves that are in different things like gases, liquids and solids [6]. But in general, acoustics is related to sound, vibration, etc. A person or a scientist who works in the area of acoustical fields is known as acoustician. The study of sounds, their frequency and the behavior of sound are included in acoustics. Acoustics is the science of production, control, transmission, response and the effects of sound. The classification of an acoustic scene allows devices to understand the environment and opens various applications [7]. For example, devices such as androids, iPhones, Internet devices, wearable devices, and robots prepared using artificial intelligence can benefit from the situations of the classification of the acoustic scene. Also, intelligence assistants represent another field that can benefit from the classification of sound scenes. IPA are software providers that make advice and perform action by automatically identifying different types of input data including audio, images, user input, context-based information, such as location, weather and private schedules. Now, Microsoft's Cortana and Apple's Siri are using audio inputs and the use of context-based information gathered from environmental sounds has a significant potential to recommend appropriate actions to users.

Environmental sound is a combination of several sound Sources. It has a lot of information that can help humans for feeling of environment around. Voice evaluation draws researchers' attention to the machine Learning has been implemented in the community and monitoring, information services on robotic navigation and tourism, etc. The recognition of acoustic events is aimed at the identification of voices. Classification is the detection of these sounds. This will helpful in information retrieval, with multimedia applications content analysis, context-aware and audio-based devices surveillance and monitoring systems. The classifications of acoustic events have aimed to the divisions of different acoustic events in target groups for specified area [8].

Fig. 1 gives an example of an acoustic classification system. Acoustic classification consists of different phases. The audio data can be classified using unsupervised or supervised learning based on data. This can be achieved by implementing different kinds of deep learning models like neural networks and feature extraction techniques to get the

features from that audio data. Afterwards, the performance of the model is evaluated for acoustic scene classification. Acoustic models are not only relating with sound classification it is also use for image classifications and some other kinds of classifications [9].

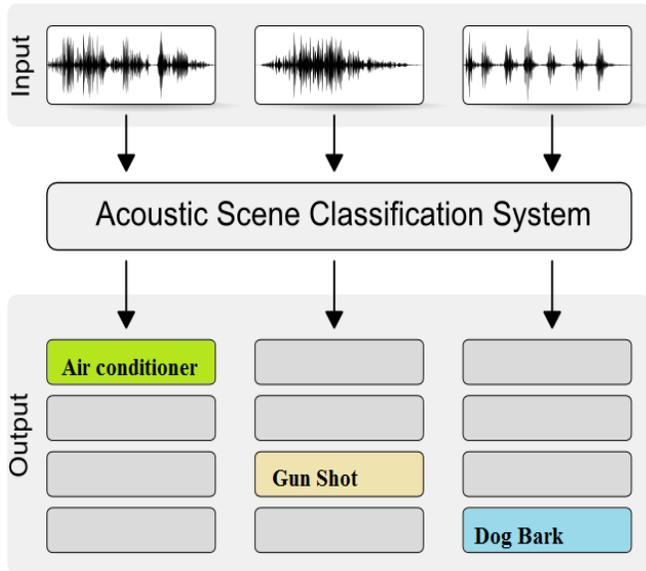


Fig. 1. Acoustic scene classification system.

B. Artificial Intelligence and Machine Learning

According to John McCarthy, artificial intelligence is the science and engineering to create the tools that are based on the sharp behavior [10]. Artificial intelligence is an intelligent behavior to make computer or software run by robots. Artificial intelligence perceives how peoples learn, how make decisions and efforts when struggling to solve the problems. The development of artificial intelligence began with the purpose of making the intelligence as same as found in the peoples. The aim of artificial intelligence is to develop the special machines (Machines that show the sharp behavior to learn, indicate, describe and advise who use it) and to perform the intelligence of peoples to systems (Develop systems which include, perceive, understand and act according to peoples). Simply Artificial Intelligence is the technique or method in which human transfer their intelligence to machines and make them intelligent to perform tasks intelligently for given data. Artificial machines get intelligence from the line of codes and also act according to the behavior of human's commands.

Artificial intelligence is scientific and technological area which has applications in other fields like Computers, Biology, Psychology, Linguistics, Mathematics and Engineering. One of the main impulses of artificial intelligence is the creation of functions for computer that are linked with the sharpness of peoples such as logical judgment, understanding and searching to solve problems. In the first half of the 20th century, people thought of artificial intelligence (AI) direct connection with the robot. Over the decades, growth in robotics proved to be enough today we have robots around us [11], but they do not stop developing at ANN.

Machine learning is an area of artificial intelligence based on the idea to make machines accessible for data and learn

themselves from the accessed data. It was started of graph identification and the idea from which the systems can be studied using no program without performing certain jobs. The peoples who want to make research they specify their interest in this field want to see computers can learn from data. From initial point of learning of machines is useful due to new representations are related to models. These systems absorb from the past adaptation to make well-grounded reflection, conclusions and outcomes. This is a discipline which is not starting except who have attained freshness. However, a variety of system studying procedures is extended over a period, the capacity to implement complex numerical solutions automatically on large amount of data.

Machine learning workflow can be described through the following figure, in which completely describe how machine learning work in a well-mannered workflow. As machine learning wants to tackle huge and most complicated problems, the issue of concentrating on the most appropriate data in a conceivable overpowering measure of information. It has turned out to be progressively critical. For instance, information mining of organizations or researcher's records regularly includes managing numerous highlights and numerous cases, the Web and the Internet have put an expansive volume of low quality data in the simple access to a learning framework. Comparative issues emerge in the personalization of the separating frameworks for data recovery, email, arrange news and so forth.

Fig. 2 depicts how machine learning algorithms works. These algorithms take a dataset as input. First, it takes a dataset to extract the features from that data using different kinds of methods of features extraction. After feature extraction use the machine learning algorithms especially deep learning techniques to grouping the data objects. At the end predict the model by taking a testing dataset, it checks the performance on that test data and at the end get the results.

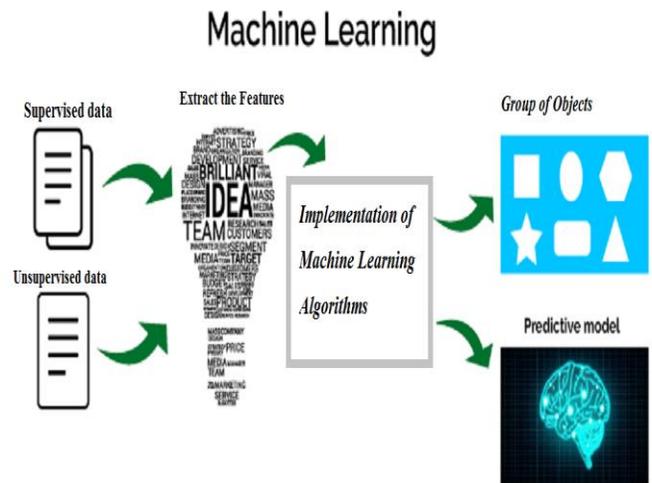


Fig. 2. Example of work-flow for machine learning.

C. Deep Learning

Deep learning is a branch of machine learning, in deep learning computational models comprising of numerous

handling layers and various reflection levels. These techniques have fundamentally enhanced the most progressive innovation in discourse acknowledgment, visual question acknowledgment, protest identification, medication revelation and genomics. Profound taking in many sided quality of expansive informational indexes, utilizing the back propagation calculation and a machine the interior parameters used to figure the impact on each layer from previous layer. Profound overlay systems have gotten leaps forward the territories of picture, video, discourse and picture handling. Voice and monotonous nets shed light on progressive articulations for example content and discourse.

Besides improving the authenticity of different patterns identification issues, one of Deep's core objectives learning, automatic machine learning revelation of numerous levels of impressions. The goal of utilizing crude information (e.g. picture pixels) as contribution to the models and let the models learn middle of the road introductions. That enables the model to learn highlight identifiers. This is particularly obvious it is imperative as demonstrated by Bengio for territories where highlights exist. It is difficult to formalize things like question acknowledgment and discourse acknowledgment errands. In the assignment of arranging woodland species, a few elective element extractors have been utilized (as noted above) demonstrates the trouble of finding a decent portrayal for inconveniences.

Recent developments in deep learning have led to a significant improvement in automatic speech recognition and music characterization. However, speech is one of many kinds of voices and people often count a variety of environmental voices to improve perception, when they are in danger and someone is walking through a busy street. It is a useful way to complement the visible information, such as more audio, videos and pictures, with the advantages that the sound can accumulate and be stored more easily. Deep learning has been a major practical success and a profound influence on machine learning and artificial intelligence literature. Practical success and deep learning in the literature have been investigated together with their theoretical features. Deep learning explores a wide range of areas for researchers to work with and learning outcomes in machine learning and artificial intelligence. Over the past decade, a number of volume control techniques have been proposed and deep learning is probably the most encouraging. As demonstrated by the deep learning, the method uses a high level representation of low level data by stacking multiple levels using a nonlinear module. Presently, deep learning architects have various variants and the convolutional neural network method is a profound learning method traditionally used for image segmentation because of its good performance in learning distinguishing local features.

Fig. 3 illustrates the layers of a neural network model. In deep learning, when it is required to learning deeply use different neural networks to classify things. These algorithms are referred to as artificial neural networks (ANNs) because the earliest use of artificial intelligence algorithms shows how learning is done in the causal mind [12]. Use different deep neural networks that can be applied for image, audio or video, speech, text, multimodal and IOT data recognition and classification while learning in depth. The deep neural network

includes an input layer that contain input nodes, hidden layers with multiple hidden nodes and an output layer that contain output nodes. Initially, set the input values at the input nodes and these nodes calculate the output values by multiplying the weights and the input values which is the same for the other hidden nodes of the hidden layers and goes to the output layers. Output layer is the last layer containing output values.

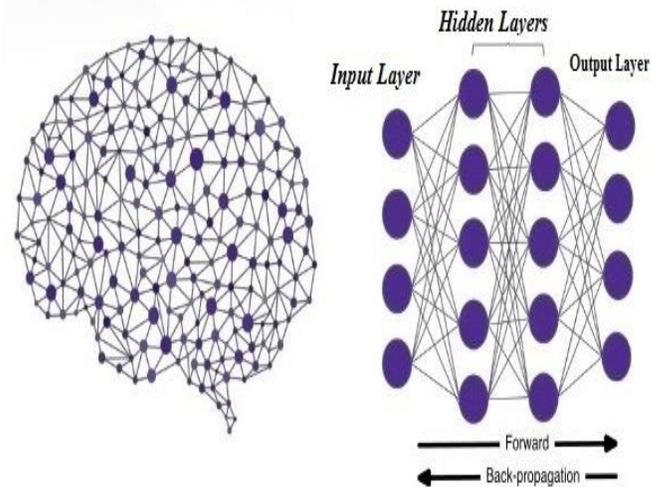


Fig. 3. Example of deep learning network [13].

II. NEURAL NETWORKS

There are different kinds of neural networks that are used in deep learning for the different kinds of working, analysis of working and to perform certain kinds of artificial intelligence application in programming way. The neural networks that are included in deep learning are Convolution Neural Network (CNN), Artificial Neural Network (ANN), Long short term memory (LSTM) and Recurrent Neural Network (RNN).

A. DNN

Deep neural network (DNN) is supervised learning feedforward artificial network used in various applications like in image processing, in video recognition, automatic speech recognition and also it is trained for acoustic scene classification. It has different layers usually an input layer, several hidden layers to build a deep architecture and an output layer. Take a DNN and train it on the data taken from DCASE 2017 challenge that contain recording of different audio scenes [14]. Input layer of DNN receives the input from data and then work according to feed forward technique and pass it to its next hidden layer. In hidden layers complete its training on the provided data set then at the end find its results according to its training on training dataset. At the end to check its efficiency implants DNN on a test data and check its working according to its testing how it works? How it classifies the data?

B. CNN

CNN is also architecture of deep learning that is used for the classification of objects on the bases of layers. It also contained layers that are named as one input layer, several hidden layers, one output layer, working of CNN is same like that DNN take input from dataset and then apply functions on it at the hidden layers and find out result and show at the output layer. Commonly used CNN layers are Max pooling layer,

convolution layer and fully connected layer. In layer that is convolution, filter is convolved with input features. Max pooling layer do the job of down sampling the input and fully connected layer connects all neurons from previous layer with its every neuron.

Fig. 4 depicts the CNN model. First, the model takes the channel input. Then this input was convoluted at next stage using 1D convolution then at the next stage Max-pooling was performed after max pooling again convoluted then fully connected layer used on that data and at the end softmax is applied on that channel input at the last stage of CNN model.

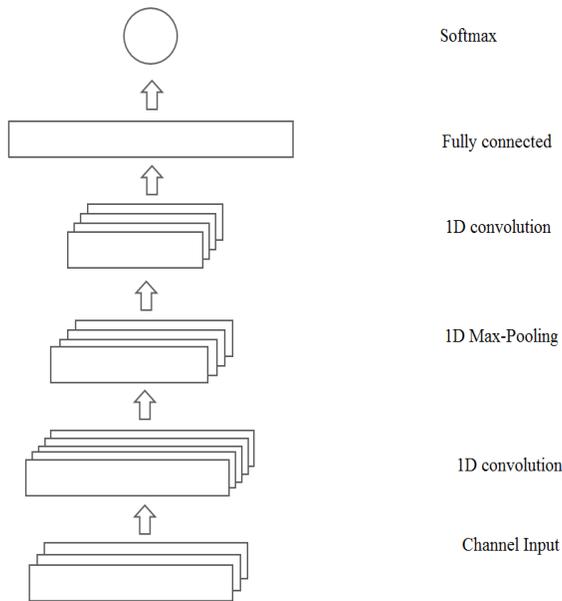


Fig. 4. CNN model.

III. AUDIO FRAMEWORK

The audio framework used in this study consists of seventeen low level descriptors for temporal and spectral sound characteristics. This set of descriptors is generic to allow a wide range of applications to use these descriptors. That can be divided into six groups (Fig. 5). In addition to these six groups, there is a very simple audio framework mute descriptor. The two basic audio descriptors are the scaled values temporarily sampled. The audio waveform descriptor describes the audio waveform envelope by sampling the maximum and minimum values in an analysis window. The audio power descriptor describes the instantaneous power levelled temporarily. These can provide a quick and effective summary of a signal, especially for visualization purposes. The four basic spectral audio descriptors have the central link deriving from an analysis of the audio signal at a temporal frequency.

The audio spectrum envelope describes the short term power spectrum of an audio signal. It can be used to show a spectrogram and to synthesize a raw naturalization of the data or as a generic descriptor for research and comparison. The center of the audio spectrum is defined as the frequency center of the power weighted record. The power spectrum can be dominated by low or high frequencies [16]. The audio

spectrum extension defines the sound moment of the power spectrum of the logarithmic frequency.

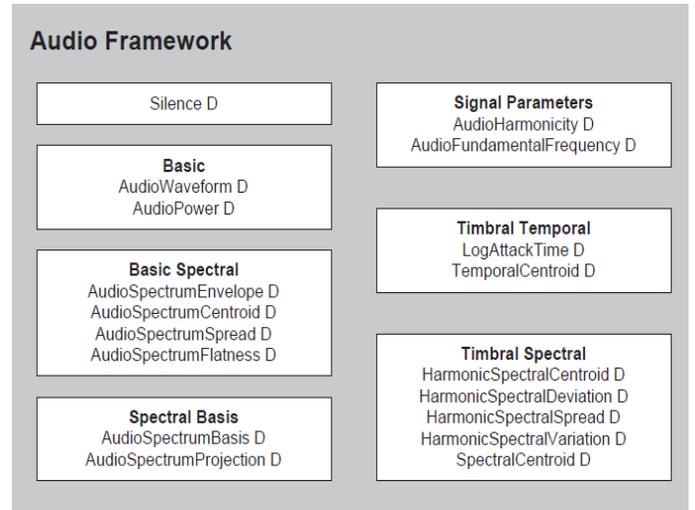


Fig. 5. Audio framework [15].

The logarithmic frequency indicates that the power spectrum is close to the centrifuge or it extends in the spectrum. The level of the audio spectrum describes the flatness characteristics of the short time power spectrum of an audio signal. This identifier refers to the scattering of the signal power spectrum from a smooth form. The spectral flatness analysis is calculated for the desired number of frequency bands. Sound signals can be used as a feature vector for robust pairing between pairs. The two spectral base identifiers represent small sized projections of a high dimensional spectral area to facilitate compactness and recognition.

The basic sound spectrum descriptor contains the basic functions used to transform high dimensional spectrum definitions to a low dimensional representation. The sound spectrum projection descriptor represents the low dimensional properties of a spectrum that is computed after projection on a reduced basis given by the sound spectrum base. The fundamental frequency is a good indication of music tone and vocal tone. Audio describes the harmonic level of an audio signal with a harmonic identifier. This makes it possible to distinguish between sounds with a harmonic spectrum and a spectrum with a non-harmonic spectrum between musical sounds and sound. The two timbral temporal identifiers describe the temporal properties of the audio segments. These are useful for describing the musical temperament. The temporal centric descriptor defines where the energy is produced over time depending on the length of the sound track. The five spectral identifiers of the timbre define the temporal characteristics of the sounds in the linear frequency domain. This makes it useful to explain the timbral spectrum coupled with the temporal identifiers especially the musical instrument tone. The spectral center identifier is very similar to the center of the sound spectrum with the use of a linear power spectrum as the only difference between them.

IV. METHODOLOGY

The methodology adopted for this research is similar to previous approaches that were based on machine learning

techniques. The most common machine learning technique used for classification of sounds or some other classifications is deep learning. In this approach, first add the dataset that contains the audio files. Afterwards, features from the dataset are extracted using the feature extraction. Then use these values in neural network layers to find out the results. The deep learning techniques are used for the classification of acoustic scenes, sounds, images or any other objects.

Fig. 6 elaborates the methodology adopted in this work. This research work uses a model that is based on two neural networks one is NN and second is CNN and implement separately. First of all, in these neural networks set the layers in three ways the layer that is at first end is called input layer that contain the input values next hidden layers that receive the input values and weights and then perform the function of neural networks and at the last end last layer that is called output layer and at that layer output values can be showed in the form of results. Secondly, use a dataset that contain the input in form of sounds use this dataset in the designed model and extract features from that dataset to take values in numeric form for the input and weight setting. To extract features, use the hybrid methods of feature extraction. Then feature extraction uses these values for finding the results.

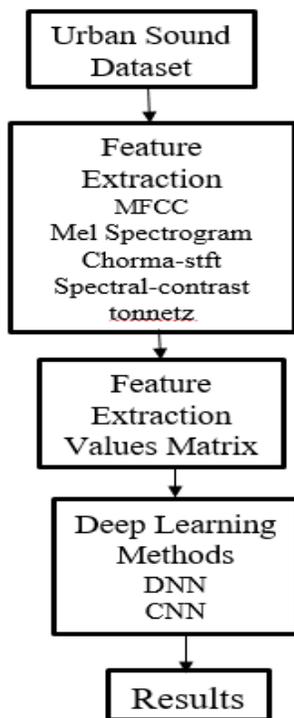


Fig. 6. Methodology adopted for this study.

A. Data Set

The dataset that used in this acoustic scene classification process was taken from the Urban sound datasets and this dataset named as Urbansound8K dataset. This dataset is divided into 10 folds and all the folds contain the audio files in the format of .wav of different kinds of sounds. These are fold1, fold2, fold3 and so on up to fold10. The dataset contains the 8732 named sound audio files and these all sounds are of

length less than four seconds. The dataset consists of urban sounds including 10 types of low-level scientific classification, i.e., air conditioning, car horn, children playing, barking dogs, drilling, idling engine, blow, jackhammer, siren and street music. To avoid big contrasts of class of property, it establishes a confinement point of 1000 denominations per class, generating a total of 8732 denominations. Table I provides details of the features used in this study.

TABLE I. SOUNDS IN URBAN SOUND DATASET

Sr. No.	Sound	Format	Duration
1	Air conditioner	.wav	<=4 sec
2	Car horn	.wav	<=4 sec
3	Children playing	.wav	<=4 sec
4	Dog bark	.wav	<=4 sec
5	Drilling	.wav	<=4 sec
6	Engine idling	.wav	<=4 sec
7	Gunshot	.wav	<=4 sec
8	Jackhammer	.wav	<=4 sec
9	Siren	.wav	<=4 sec
10	Street music	.wav	<=4 sec

B. Feature Extraction

To extract the useful features from sound information, utilize the Librosa library. It gives a few strategies to extract diverse features from the sound files. Following are strategies to extract different highlights:

- **Spectrogram Mels:** calculates an energy spectrogram on the Mel scale
- **mfcc:** coefficients of cepstral Mel frequency
- **Chorma Stft:** calculates a chromatogram from a waveform or a power spectrogram
- **Spectral Contrast:** calculates the spectral contrast
- **Tonnetz:** calculate the characteristics of the tonal centroid (tonnetz)

Fig. 7 illustrates that how to simplify the procedure for highlighting the extraction of audio locks, two assistance strategies have been identified. First, parse audio files which takes the name of the main catalog, the subdirectories within the main index and the expansion of the document (the default is .wav) as information. At that time, it emphasizes each of the documents within the subdirectories and calls the second job called the associated function. Take the document as information, read the record by calling the librosa, load technique, concentrate and return the salient points mentioned above. In general, these two strategies are necessary to move from the raw sound clicks to the salient points of the instructions. The class name of each solid closure is in the document name.

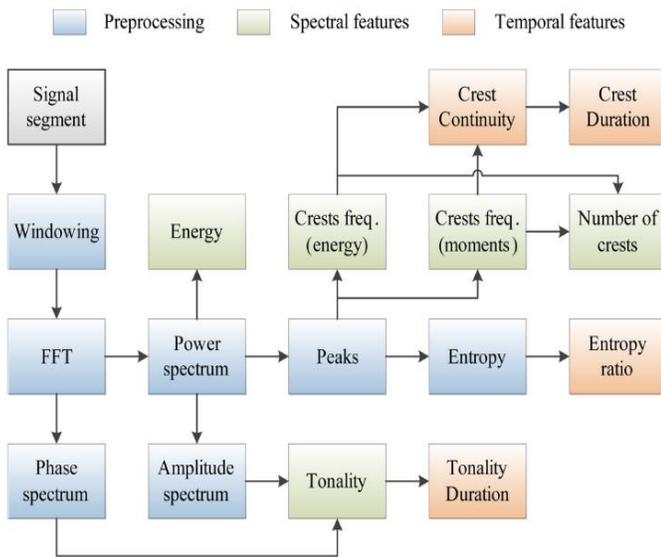


Fig. 7. Feature extraction from audio sounds.

The channel bank has a generally trapezoidal size reaction, with a significant portion between adjacent channels. In each channel, a flag envelope setting shown significantly with changing wave and low pass screening. Each flag is standardized on the calculation of the normal level on a complete articulation and isolate this quantity. Moderate adjustments in each standardized dissected indicator envelope then sieving flag through an unpredictable band passage channel and taking the log expansion yield.

C. Required Tools

The tools that are used for this acoustic scene classification process of sounds are as following:

- Anaconda
- Spyder
- Python

D. Required Libraries

Libraries that are used in this research work as follows:

- Numpy
- Scipy
- Pandas
- Matplotlib
- Plotly
- Theano
- Tensorflow
- Keras
- Librosa
- FFMPEG

V. RESULTS AND DISCUSSION

Two neural networks are implemented in this research work with different feature extraction techniques to get the results over the Urban dataset 8k. Extract the features from the dataset that putted in the working code after feature extraction set the hidden layers and weights on those layers by using the feature values that extracted from the data. Data set contain the sounds of each classes and in all the folders sounds of all classes are presents so selection of folds is very sensitive matter. The two neural networks that are used in this research work are DNN and CNN.

In deep neural network use the three folds of dataset and then parse that sound files and parsing that sound files extract the features from the sound files. By extracting features use the values of that features in the form weights on hidden layers and then apply the formula of DeepNN and get the results. Fig. 8 depicts the F-score obtained from that DeepNN. Approximately 80% accuracy was achieved using this approach.

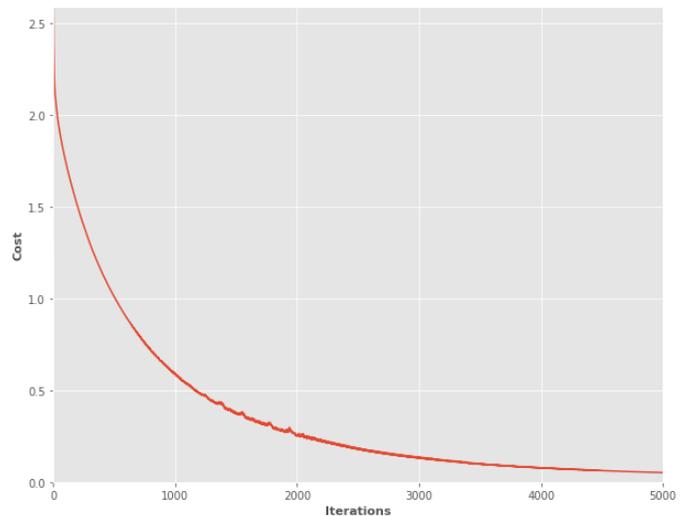


Fig. 8. Resulting curve of DNN.

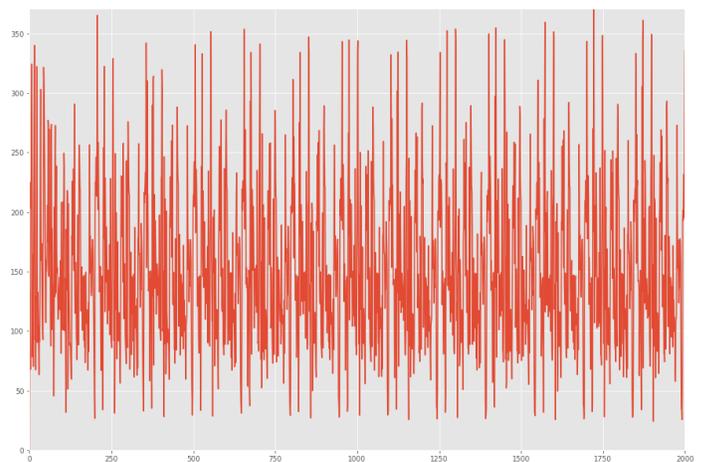


Fig. 9. Resulting graph of CNN.

The mfcc technique was used to extract the features from the sound files, two folds of dataset, 41 frames, 60 bands, 2460 feature size, 10 class labels, 2 channels, 50 batch size, 30 kernel size, 20 depths, 150 hidden units, 0.01 learning rate and 2000 training iterations. In CNN also weights are set by using the values of feature extraction that are extracted using the librosa library. Fig. 9 depicts the resulting plot that is obtained from the CNN.

Accuracy obtained from the CNN is 0.153 nearly about to 87% that is enhanced from the previous approach. In CNN working is slower but the results are accurate when compared to the DeepNN.

VI. CONCLUSION

The results after implementation of different hybrid feature extraction techniques on Urban dataset 8K improved in both machine learning techniques CNN and DNN. The fundamental task of this research work was that how improve the results from the previous methodology that is adopted on that Urban sound dataset 8K, so use many feature extraction methods to extract the feature values from sound data then apply deep learning methods DNN and CNN to get the results from extracted features. After implementation of this hybrid feature extraction method improve the efficiency of the neural networks on the described dataset.

REFERENCES

- [1] Barker, J., Vincent, E., Ma, N., Christensen, H., & Green, P. (2013). The PASCAL Chime speech separation and recognition challenge. *Computer Speech & Language*, 27(3), 621-633.
- [2] Benetos, E., & Dixon, S. (2013). Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *The Journal of the Acoustical Society of America*, 133(3), 1727-1741.
- [3] Ramona, M., & Peters, G. (2013, May). Audio Print: An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In *Acoustics, Speech and Signal Processing (ICASSP)*, 2013 IEEE International Conference on (pp. 818-822). IEEE.
- [4] Marques, P. A., & De Araujo, C. B. (2014, October). The need to document and preserve natural Soundscape recordings as acoustic memories. In *Proceedings Invisible Places Conference*. Draft version available: <http://invisibleplaces.org/pdf/ip2014-marques.pdf>. Accessed (Vol. 13).
- [5] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10), 1733-1746.
- [6] Stowell, D., & Plumbley, M. D. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2, e488.
- [7] Bisot, V., Serizel, R., Essid, S., & Richard, G. (2016, March). Acoustic scene classification with matrix factorization for unsupervised feature learning. In *Acoustics, Speech and Signal Processing (ICASSP)*, 2016 IEEE International Conference on (pp. 6445-6449). IEEE.
- [8] Boudreaux, K. (2018, March). Commonsense and Artificial Intelligence Systems. In *Society for Information Technology & Teacher Education International Conference* (pp. 21-24). Association for the Advancement of Computing in Education (AACE).
- [9] Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32(3), 16-34.
- [10] Mesaros, A., Heittola, T., & Virtanen, T. (2016, August). TUT database for acoustic scene classification and sound event detection. In *Signal Processing Conference (EUSIPCO)*, 2016 24th European (pp. 1128-1132). IEEE.
- [11] Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013, September). A database and challenge for acoustic scene classification and event detection. In *Signal Processing Conference (EUSIPCO)*, 2013 Proceedings of the 21st European (pp. 1-5). IEEE.
- [12] Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited.
- [13] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.
- [14] Hovy, E., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194, 2-27.
- [15] Andersson, Tobias. *Audio classification and content description*. 2004.
- [16] Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Quality Assurance for Data Analytics

Rakesh Kumar, Birth Subhash, Maria Fatima, Waqas Mahmood
Faculty of Computer Science
Institute of Business Administration
Karachi, Pakistan

Abstract—Quality Assurance is a technique for ensuring the overall software quality suggested by Global Standards bodies like IEEE. The Quality Assurance for Data Analytics requires more time and a very different set of skills because Software Products, which are used for Data Analytics, are different than that of traditional ones. In result, these Software Products require more complex algorithms to operate and then for ensuring their quality, one needs more advanced techniques for handling these Software Products. According to our survey, Data Analytical Software Products require more work because of their more complex nature. One of the possible reasons can be the volume and variety of Data. On the same hand, this research emphasizes on testing of Data Analytical Software Products which have many issues because testing of these Software Products requires real data. However, every time the testing of these Software Products is based either on dummy data or simulations and these Software Products fail when they work in real time. For making these Software Products work well before and after deployment, we have to define certain Quality standards. In this way, we can get better result producing analytics Software Products for better results.

Keywords—Software Quality Assurance (SQA); data analytical softwares; data driven softwares; real time analytics; data analytics; quality issues; quality control

I. INTRODUCTION

Quality Assurance (QA) related activities are proven to be beneficial because it gives a confidence about the completion of requirements and needs that a user expect from a system, it can be the quality of product or it may be the accessibility and reliability, accessibility and reliability are just two qualities of systems, there are many more suggested by different experts. With change in time and access of internet, everything has changed its nature especially the field of information technology with increased usage of data driven softwares has impacted a lot. The more softwares will rely on data the more we will be facing challenges of software quality [1], [2].

In this paper we will not only discuss the quality issues in data driven softwares but we will also discuss best practices for testing of data analytics. Data analytics has contributed in almost every sector, especially for making life easier and for the betterment of our society, whether it is health sector or smart homes, data analytics has a huge impact on human race [3].

The increase in importance of data analytics has increased the work for developers especially for quality assurance teams because as we know and we have also practiced that Quality Assurance activities are the major portion of the effort in development of a software product. We have seen tremendous

change in data related work and as a result the complexities and quality issues in software products have grown since few decades, these gaps in Software development must be handled with the help of performing activities to perform software quality assurance. Hence there is a need to define certain measures and approaches that must be followed to tackle quality issues of data driven softwares or data analytical softwares [4], [5].

Proliferation in the volume and variety of data has challenged practitioners to work more on the authenticity of data [6]. Let's discuss the internal structure and creation of data driven or data Analytical softwares because before talking about Quality Assurance we must know about the structure, development and purpose of those systems. Main goals of Data Analytical products are to read the data, understand the data and then find trends from it and most importantly to predict about the subject matter. These predictions can be pointing to a grater meaning of subject that is being observed, and results obtaining from these analysis can be leading to new theories and innovations, that is why these analytical softwares are getting familiarity day by day.

These softwares generally work in three stages: Data Preparation, Data understanding, and Prediction. This whole process continues after data gathering; data gathering can be done by doing surveys, focus groups or by taking interviews, etc.

- Data preparation that includes ETL (Extract, Transform and load), ETL includes data cleaning activities like, normalization, formatting the data into specified format and removing redundancies etc. Many times we need to make categories of continuous data.
- Data Understanding can be referred as Understanding of data by software, First software understands the data and then it is in the state to predict something by applying various predictive models.
- Prediction is the major part of any Analytical Software; it can be done after understanding of data by software and then applying some analytical method on it.

Quality Assurance is not only required for software development but it has many other benefits as well, software quality assurance also includes the contractual condition that are very essential when someone (an individual or an organization) is outsourcing a software; Q.A suggests best practices, rules and many other things like budget and deadlines to decide before signing the agreement [7].

This research focuses on the complexities in data analytical softwares and suggestions for quality enhancement in softwares [8].

II. BACKGROUND

Availability of internet at every place especially in industries has a huge impact on the production of data, earlier data used to be stored in disks or hard drives now trend of storing data has moved towards digitization because data is producing in streams every second. Especially in the field of health where data is everything and to handle and manage this huge amount of data we need some special software that is totally different from traditional softwares. To develop tools/softwares that are data driven or we can say that are used for analysis purpose, we must define certain quality attributes because as we have already mentioned that these analytical softwares are very different and complex from traditional softwares [9].

Researchers from past have practiced many approaches to gather data and relevant information about the Research topic. Many methodologies and strategies to conduct a survey have also been proposed by practitioners and they have stressed the adaptation of various evidences related to research area [10]. With technological advancement and usage of social networking, the Data generating apps have introduced us with the big data and these apps generate such a huge amount of data that can be estimated as Terabytes, in a single day worldwide. Medical field also needs a quality assurance mechanism for their figure data as most of the clinical data contains medical figures, such as ex-rays. On an average a medium size hospital generates about 1 million figures per year [11]. To get some results from that data and its storage is not an easy task that's why we need some special techniques and standards that can tackle this issue [12].

This research will be focusing on the data analytical softwares' quality measures as we have seen changes in trend of softwares. This domain is still new and needs certain rules and standards so that quality of analytical software can be measured. We will be having a research survey along with a literature review, because research in this domain is not yet properly done so we will have to conduct a research survey too. The research survey will consists of questionnaire related to complexities and practices used in industries, this survey will be conducted from industry persons and students who have at least a bachelor degree in computer science because they are the one who will be working for those analytical softwares. Industry persons will also be taken interviews so that we can have a better knowledge about our research, since this is still a developing domain we will try to take surveys and interviews from people living outside of Pakistan.

III. LITERATURE REVIEW

Software Quality Assurance is a methodology that suggests globally accepted practices and standards to assure software quality. It aims to provide a quality product by conducting tests at different stages of Software project development [13].

Quality Assurance for data analysis needs domain knowledge in the distinctions of not only what can be the

complexities, errors and efforts can be required at the time of collection of data, but interpretation of data can also be a huge challenge, because results from data can be misleading many times. That is why it is highly recommended that Quality Assurance team must be involved in the development phase and they should get in touch directly to the developers. These approaches can be followed when working on a project related to data Analytics:

Make a Quality Assurance group, who will verify that the output data that Data Analytical system produces is valid and give outcomes or results that we are supposed to get. Working with experienced individuals is necessary when we are working on some analytical system, because results can be dangerous many times. Adding Quality Assurance persons in developing team can be beneficial because they will be finding and resolving quality related issues at earliest and this approach can save out time, as a result, money can be saved. It is a mandatory task to plan for quality assurance activities, Quality Assurance team must make plans for testing; test plans can be interpreted as test cases etc. This approach gives a view to Quality assurance team and with this approach we can see a positive movement in our project testing side [4].

Testing in real environment is complex and money wasting technique, that's why to simulate devices is a better option and this technique is successful so far, but in case of data analytics we need real data, like in case of traffic analysis we cannot simulate it or visualize it, though we can but it is not a successful technique. Recently Automated vehicle crash has occurred and this minor accident has caused company so much loss. In result, for data analytics we need real and variety of data so that beneficial output can be generated [14].

These approaches are very necessary because in majority of the systems data sources are too many and data is too huge to check it manually. For a simple query for example how many bike crossed this signal at 7:00 pm yesterday? Now this seems very simple in first sight but when we will be running this query manually it will take many hours may be a day or two in traditional or manual systems, that's why we need to migrate to especial systems and we need to concentrate on quality more [15].

Many researches have highlighted the software quality assurance as a recent and autonomous field and it is also mentioned that software quality assurance is introduced after hardware quality assurance and similarities between these two domains are also discussed. Traditional methods are not enough to do the necessary improvements; there need to do some changes [16].

The purpose of this project is to highlight such quality attributes and techniques that must be present and used when developing or testing analytical software.

IV. METHODOLOGY

This research paper is based on literature review and questionnaire survey; we have conducted a research survey from variety of practitioners, those are either industry experts or students who have fresh and innovative ideas related to this domain. Motivation behind this survey was the unavailability of enough literature that could support our research [17].

V. SURVEY RESPONSES

We have conducted a survey on Quality assurance for Data analytics. Through in this survey we have found so many different views. Some people think quality assurance for data analytics is really important while other thinks it is not important. Many people who have taken part in the survey are either employees or final year students. This summary will be showing the results from our survey.

We wanted to make sure that our audience must have CS background or at least they have enough knowledge about quality assurance so we asked them about their professions. People, who have given their views on this survey, are mostly students and developers. The graph summary in Fig. 1 showed that almost 60% are students, 30% are developers and remaining 10% are Instructors, Business Analyst and SAP consultants. Most students are graduates who have experience in this particular field. During this survey, we have analyzed the importance of quality assurance in every profession.

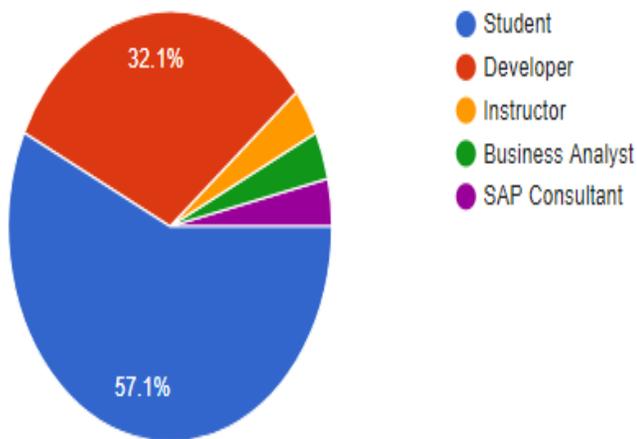


Fig. 1. Profession.

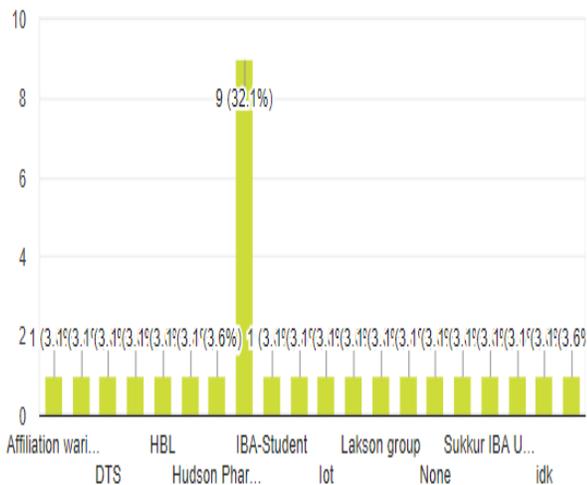


Fig. 2. Affiliations.

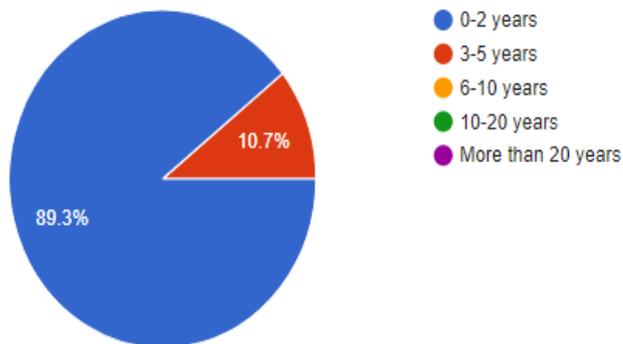


Fig. 3. Experience in years.

We also wanted to observe the people’s affiliation; in which organization they are working or in which organizations they have previously worked. Fig. 2 shows that most people are either IBA students or they work in Hudson pharma Pvt. People from cloudKibo, Lot, HBL, Lakson group and Sukkur IBA have also helped us to analyze the need of quality assurance for data analytics. We can observe in this figure, the affiliation is on horizontal line and no: of employees on vertical line.

For this survey, it was very important for us to know about the work experience of the people who are taking part in this research, because mostly experienced people support quality assurance than fresh practitioners. In Fig. 3, we can see that most people have either no experience or less than three year experience. Graph summary shows that, only 12% are those who have 3 to 5 year experience and all remaining have 0 to 2 year experience. We tried to reach out to those employees who have more than six year experience but unfortunately we were only able to get 2 or 3 responses.

We wanted to know if people think that quality assurance is important for any software product. As it has already mentioned that most people who have taken part in this survey, are either final year students or employees because they have better idea of quality assurance rather than students who are in their first, second or third year. In Fig. 4 we can see, 96.2% of them think that it’s really important and only 3.8% people think quality assurance is not much important. The reason behind this is maybe they don’t have too much experience or maybe they are over confident on their code.

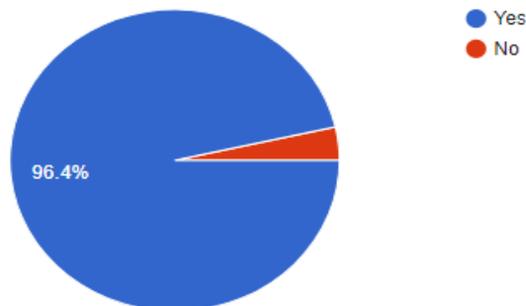


Fig. 4. According to you, is quality assurance important for any software product?

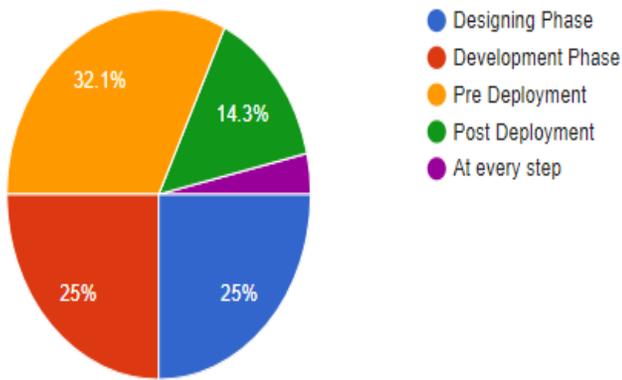


Fig. 5. At which stage of software development, quality assurance team should interfere?

We wanted to analyze, at which point quality assurance team should play its part. From Fig. 5, we can observe that everyone has its own view. 30.8% of them suggested that QA team should play its part before the deployment only, 15.4% of them suggested that they should apply tests and algorithms after deployment, 26.9% of them suggested that QA team should be there for only development phase and very few of them suggested that QA team should take part at every step and also after the deployment as well.

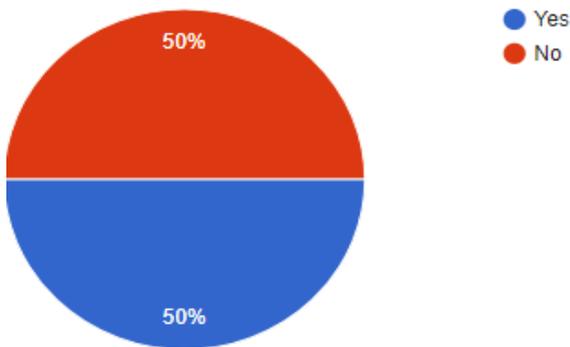


Fig. 6. Have you ever worked in the development of data analytical software.

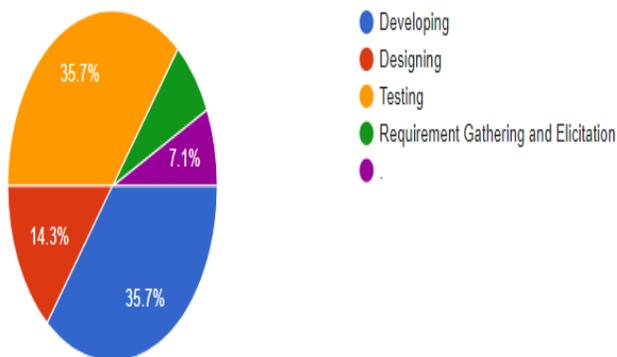


Fig. 7. If yes, what was your role?

We are conducting survey for data analytical software products that is why we wanted to observe our users in detail. Fig. 6, shows the responses of a question in which audience were asked, weather they have ever worked on any data analytical software. The summary told us that 50% of them have previously worked on data analytical software and they also believe that Quality assurance for Data analytics is really important.

From Fig. 7, we can understand that results are very interesting because question is answered by only those individuals who have ever worked for data analytical software and it is very important to know the views from them who have ever worked for an analytical software, almost 1/3 of the population have worked in the development of analytical software products and almost same number of Practitioners have worked in testing of those softwares, around 1/7 of the sample has worked in the designing phase of that software while a few people have worked in the requirement gathering phase of any analytical software ; requirement gathering can be included in designing phase.

From these results we can claim that almost half of the topic surveyors have worked in the creation of data driven or data analytical software products.

Fig. 8 gives us an overview about the opinion of people regarding the complexities of data driven or data analytical softwares.

The purpose of this question was to understand and find what people think about the complexities related to data analytical softwares. They were asked if complexities are different than the traditional softwares. Majority of the people that is around 4/5th of the total sample has accepted that complexities in data analytical system are more than the traditional softwares.

These results show that most of the Practitioners have clear understanding of data analytical softwares or at least they have a knowhow of these softwares. Those people who said the complexities aren't different it may be because they do not have ever worked for analytical softwares or maybe they have found it easy for them.

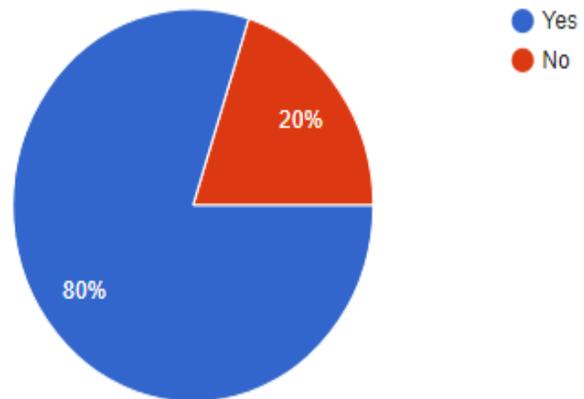


Fig. 8. In your opinion, are the complexities in data analytical softwares are different than the traditional softwares?

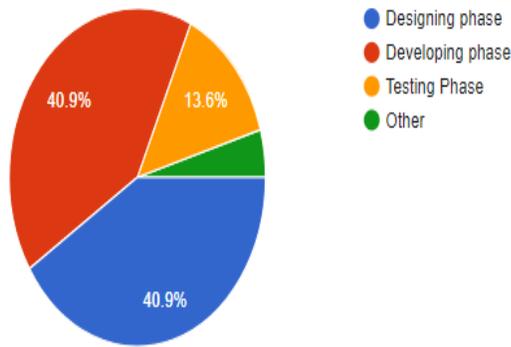


Fig. 9. If yes, in which phase these complexities occur?

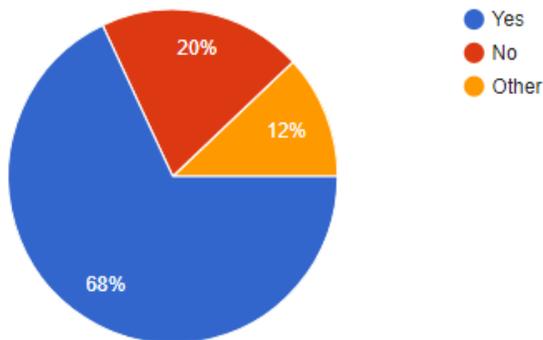


Fig. 10. Do you think there are any gaps in quality assurance of data analytical softwares?

This question was asked just to understand that during which phase practitioner find more complexities when working on analytical softwares. Fig. 9 explains, around 2/5th of the sample have said that they have found complexities during designing phase and I think this is genuine response because designing phase is the initial phase where software development team should be ready for upcoming complexities and should plan accordingly. Moreover almost same number of sample have expressed that they felt complexities during development phase that is also a genuine thing because if we do not plan and get ourselves ready during designing phase, we will having issues during development phase and almost 1/7th of sample feel that they face some complex situation during testing phase that is also understandable because testing of these softwares is also not an easy task, we need dummy data to simulate real life scenarios but many times real life results can be different than simulations , that means testing of these softwares is also a big issue. While few people also think that there can be complexities except these phases too.

Fig. 10 shows the responses of question “Do you think there are any gaps in quality Assurance of Data Analytical Softwares?” around 3/4th of the audience have said yes there are many gaps in the quality assurance of data analytical softwares, these gaps can be referred as barriers to test the analytical softwares, most important thing when testing these types of systems is data, because we have not sufficient data at the time of testing and in the end our software fails in real

environment. While 1/5th of the population thinks that there isn't any gap in the quality Assurance of these Analytical softwares it may be because they haven't tried this task with their own hand or may be because they have find it easy or it is also a fact that they have find it interesting because many people find these analytical softwares very interesting. In addition to this 1/7th people have chosen others option it may be because they haven't ever worked for these kind of softwares or maybe they have some different views about this.

Fig. 11 shows the summary of the responses, in which people were asked about the availability of the solutions related to these complexities, around 3/4th of the total population have positive opinion that there are many solutions to tackle these complexities. Although there are also few that was around 1/7th of population who do not know what to comment on this, that's may be because they haven't worked on any related software. Around 1/9th of population said no there is not any solution available that can tackle these complexities, that's may be because they have worked on more complex softwares.

Fig. 12 shows the summary of the responses of a question, in which we asked our audience, if they allow us to contact for further details, around 3/5th of population is ready to tell us more about their experience related to this survey subject, few people said no, they do not want to be contacted and while 1/5th of the sample also chose other option that can be interpreted as they have some conditions before contacting them.

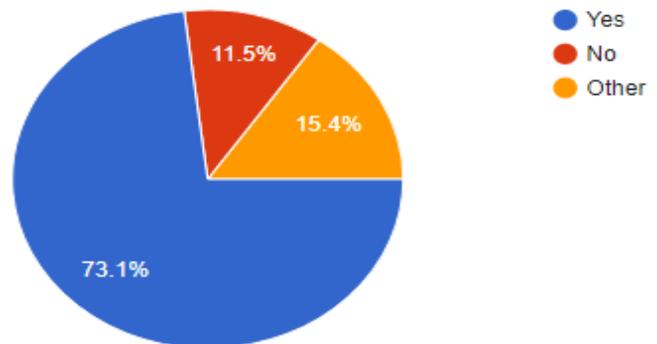


Fig. 11. Are there any solutions available to these complexities?

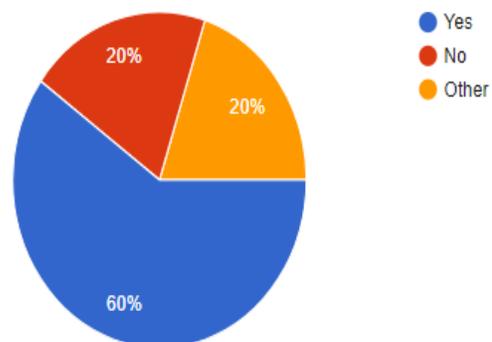


Fig. 12. For further information if we want to contact you personally, are you willing to help us in this regard?

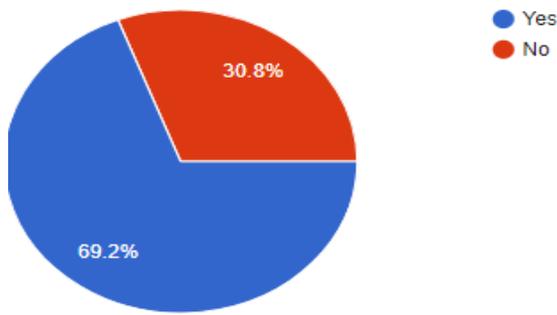


Fig. 13. Can we use your email address for contact purpose.

We also asked them if we can contact them on the same email address that they have provided in the beginning of survey form, Fig. 13 shows that most of them that was around 2/3rd of the population said Yes they can be followed on same email address, while 1/3rd of the sample doesn't want to be contacted on same email address that may be because they do not want to be contacted anymore.

This Survey gives us the idea what practitioners and concerned people think and has experienced while working on the data analytical softwares, all these responses will be helpful in our research and we appreciate each person who helped us achieve our goal.

VI. CONCLUSION

Quality Assurance for a software product is an essential part of development cycle and when it comes to Data Analytical Software, responsibility of Quality Assurance team increases due to the complexities because of such a huge amount of data and its variety. We have also conducted a survey to know about the views of practitioners those have worked/working in this domain. Majority of them have accepted that development of an Analytical software is more complex than a traditional software, and it was also answered that majority of them suffered issues during development of these software products while there are also few who suffered during testing phase. Our research concludes that there is a strict need of Quality Assurance team to interfere and suggest some solutions that can bridge that gap of quality requirements. Surveyors have also a positive feeling that these gaps can be filled by some new techniques and by applying those techniques, Quality Requirements for Data Analytical software can be fulfilled.

REFERENCE

[1] M. Bruneforth, Martin and I. V. Mullis, Quality Assurance in Data Collection, vol. 10, M. O. Martin and I. V. Mullis, Eds., Boston: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1996.

[2] L. Stahl, "Quality Assurance Project Plan for Data Analysis Activities for the National Study of Chemical Residues in Lake Fish Tissue," U.S. Environmental Protection Agency, Office of Science and Technology, Washington, D.C. 20460, 2007.

[3] F. Diko, Z. Alzoabi and m. Alnoukari, "Enhancing Education Quality Assurance Using Data Mining," 2016.

[4] "Quality Assurance for Analytics: 4 Steps to Avoid Big Headaches," 22 August 2017. [Online]. Available: <https://www.qualitylogic.com/2017/08/22/quality-assurance-for-analytics/>. [Accessed 04 April 2018].

[5] H. Foidl and M. Felderer, "Data Science Challenges to Improve Quality Assurance of Internet of Things Applications," in *International Symposium on Leveraging Applications of Formal Methods*, 2016.

[6] Narada Wickramage, "Quality assurance for data science: Making data science more scientific through engaging scientific method," in *Future Technologies Conference (FTC)*, San Francisco, USA, 2016.

[7] D. Galin, *Software Quality Assurance From theory to practice*, Pearson education limited, 2004.

[8] J. Sargeant, "Qualitative Research Part II: Participants, Analysis, and Quality Assurance," *Journal of Graduate Medical Education*, vol. 4, pp. 1-3, March 2012.

[9] W. Raghupath and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and System*, p. 10, 2014.

[10] R. Per and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Download PDF*, April 2009.

[11] V. S. Moustakis and L. Tsironis, "Knowledge Quality Assurance in Medical Data Mining," in *Proceedings of International Conference on Information Quality Management*, Chania, Greece, 1996.

[12] C. Tao and J. Gao, "Quality Assurance for Big Data Application— Issues, Challenges, and Needs," *National Natural Science China*, p. 7, 2009.

[13] S. Farooqui and W. Mahmood, "A survey of Pakistan's SQA Practices: a Comparative Study," in *29th International Business Information Management Association Conference*, Vienna, Austria, 2017.

[14] F. Lambert, "Tesla Autopilot confuses markings toward barrier in recreation of fatal Model X crash at exact same location," 3 April 2018. [Online]. Available: <https://electrek.co/2018/04/03/tesla-autopilot-crash-barrier-markings-fatal-model-x-accident/>. [Accessed 4 April 2018].

[15] SeattleDataGuy, "Data Quality Is Not as Sexy As Data Science," 15 September 2017. [Online]. Available: <https://medium.com/@SeattleDataGuy/good-data-quality-is-key-for-great-data-science-and-analytics-ccfa18d0ff8>. [Accessed 4 April 2018].

[16] F. J. Buckley and R. Poston, "Software Quality Assurance," *IEEE Transactions on Software Engineering*, pp. 36-41, 1984.

[17] R. M. Groves, F. J. Flower Jr., M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau, *Survey methodology*, 2 ed., John Wiley & Sons, 2009, p. 488.

ANNEXURE-A

Questionnaire:

Email address:

Profession:
Ex: Developer, Student

Affiliation:
Ex: IBA, Arpatech etc

Question: Experience in Years?

- 0-2 years
- 3-5 years
- 6-10 years
- 10-20 years
- More than 20 years

Question: According to you, Is quality Assurance important for any software Product?

- yes
- No

Question: At which stage of software development, quality assurance team should interfere?

- Designing phase
- Development phase
- Pre-deployment
- post deployment
- Other. . . .

Question: Have you ever worked in the development of Data analytical Software?

- Yes
- No

Question: If yes, what was your role?

- Developing
- Designing
- Testing
- Others

Question: In your opinion, are the complexities In Data Analytical softwares are different than the traditional softwares?

- Yes
- No

Question: If yes, in which phase these complexities occur?

- Designing Phase
- Developing Phase

- Testing Phase
- Others

Question: Do you think there are gaps in Quality Assurance of Data Analytical Softwares?

- Yes
- No
- Others

Question: Are there any solutions available to these complexities?

- Yes
- No
- Others. . . .

Question: For further information if we want to contact you personally, are you willing to help us in this regard?

- Yes
- No

Question: For further information if we want to contact you personally, are you willing to help us more in this regard?

- Yes
- No.

Question: Any suggestions or feedback?

Ans:

Developing Communication Strategy for Multi-Agent Systems with Incremental Fuzzy Model

Sam Hamzeloo, Mansoor Zolghadri Jahromi
Department of Computer Science and Engineering
Shiraz University
Shiraz, Iran

Abstract—Communication can guarantee the coordinated behavior in the multi-agent systems. However, in many real-world problems, communication may not be available at every time because of limited bandwidth, noisy environment or communication cost. In this paper, we introduce an algorithm to develop a communication strategy for cooperative multi-agent systems in which the communication is limited. This method employs a fuzzy model to estimate the benefit of communication for each possible situation. This specifies minimal communication that is necessary for successful joint behavior. An incremental method is also presented to create and tune our fuzzy model that reduces the high computational complexity of the multi-agent systems. We use several standard benchmark problems to assess the performance of our proposed method. Experimental results show that the generated communication strategy can improve the performance as well as full-communication strategy, while the agents utilize little communication.

Keywords—Multi-agent systems; decentralized partially observable Markov decision process; communication; planning under uncertainty; fuzzy inference systems

I. INTRODUCTION

One of the main goals of artificial intelligence is designing autonomous agents interacting in a domain. A Multi-Agent System (MAS) includes multiple autonomous agents operating in an uncertain environment in order to maximize their utility. In MAS, each agent independently perceives its local environment and influence the environment by executing its actions. Many artificial intelligence problems can take advantage of MAS design such as multiple mobile robots, sensor networks, disaster response teams, smart city and video games.

There are two types of problems in MASs, *self-interested* and *cooperative* settings [1]. In self-interested scenario the agents can have different and even conflicting goals. In cooperative setting, which we focus on it in this work, the agents cooperate to reach a shared target. In this case, each agent individually makes a decision based on its local observation, but the maximum reward will be achieved when the individual decisions are coordinated. Communication is an important factor to preserve coordinated behavior. However, communication is not always available, especially when the agents have limit on battery usage or the communication channel is noisy or limited. Therefore, one of the main challenges in MASs is to maintain coordination over a long period of time with minimal communication.

Various mathematical models have been used to characterize decision-making problems. In a stochastic fully observable environment, Markov Decision Process (MDP) provides a powerful modeling tool. Partially Observable Markov Decision Process (POMDP) is employed in problems with limited sensing capabilities. Decentralized Partially Observable Markov Decision Process (Dec-POMDP) is a powerful framework for collaborative multi-agent planning in an uncertain environment [2]. In this paper, we use Dec-POMDP to model cooperative MAS problems.

The strategies of multi-agent problems are categorized in two categories, finite-horizon and infinite-horizon Dec-POMDPs. Finite-horizon policies are usually represented by a decision tree and numerous techniques have been proposed to obtain or approximate the optimal policies [3]-[5]. Moreover, a number of methods have been developed to generate decentralized policies with minimal communication usage [1], [6]. On the other side, finite state controllers (FSCs) is a major model to represent infinite-horizon Dec-POMDP policy. Several optimization techniques have been used to approximate the parameters of FSCs, for example, Linear programming [7], nonlinear programming [8] and expectation-maximization [9], [10]. However, identifying best situations for communication has not been considered by existing methods. This paper focuses on solving this issue.

One of the powerful function approximators are fuzzy systems that can approximate any non-linear system to an arbitrary accuracy. A fuzzy system is capable of handling high level of uncertainty by a compact fuzzy rule-base. Therefore, presenting fuzzy model is desirable to solve a MAS in previous studies [11]. In [12] an incremental fuzzy controller has been introduced to find a solution of large MASs.

Our aim in this paper is to present an algorithm to identify best situations for making communication in MASs modelled by infinite-horizon Dec-POMDP. This method develops a strategy that helps the agents to maintain coordination with minimal communication. This communication policy is developed centralized in a training phase, where the communication is not restricted. The agents use this policy decentralized in a test environment that the communication channel is limited. This paper presents an incremental method to estimate the benefits of communication in every possible situation that the agents can have. Based on this estimation, the agents can decide when the communication has the most impact on the improvement of the final performance. The results show that the performance of the presented

communication strategy is almost the same as the full communication.

The organization of the rest of paper is as follows. Section 2 formally defines the infinite-horizon Dec-POMDP and gives an overview of the Dec-POMDP solution methods. Section 3 presents the details of the proposed method. In Section 4 we evaluate the proposed communication strategy on several well-known Dec-POMDP problems. Finally, the conclusions are given in Section 5.

II. BACKGROUND AND RELATED WORKS

In this paper, we consider a group of agents cooperate with each other in an uncertain environment over infinite time steps. At each time step t , the agents take joint action \vec{a}^t (action a_i^t for i -th agent) that causes the state of the environment to change from s^t to s^{t+1} . After that, each agent perceives its observation and receives a global reward from the environment. This cycle repeats over infinite steps. This type of MAS problems is properly modelled by infinite-horizon Dec-POMDP [12]. Fig. 1 displays the interaction of the agents and the environment.

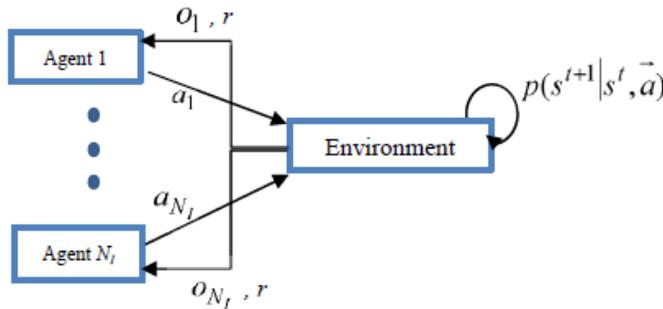


Fig. 1. Dec-POMDP Setup.

A. Infinite-Horizon Dec-POMDP

In infinite-horizon Dec-POMDP, a group of agents are considered that operate in an uncertain environment over infinite steps. Infinite-horizon Dec-POMDP is a tuple $\langle I, S, \{A_i\}, \{\Omega_i\}, P, O, R, b^0, \gamma \rangle$ where I is a finite set of agents and S is a finite set of states. Each state determines the specific situation of the environment. The number of agents is N_I and N_S is the number of states. A_i and Ω_i specify the finite set of actions and observations available for agent i . $\vec{a} = \langle a_1, \dots, a_{N_I} \rangle$ denotes a joint action ($\vec{A} = \times_{i \in I} A_i$) and $\vec{o} = \langle o_1, \dots, o_{N_I} \rangle$ denotes a joint observation ($\vec{\Omega} = \times_{i \in I} \Omega_i$). If the agents take joint action \vec{a}^t in time step t , the state of the environment is transitioned from s^t to s^{t+1} with probability $P(s^{t+1} | s^t, \vec{a}^t)$. The probability of the joint observation \vec{o}^{t+1} in state s^{t+1} after the agents perform joint action \vec{a}^t is $O(\vec{o}^{t+1} | s^{t+1}, \vec{a}^t)$. At the end of each time step, the environment gives the agents the global reward $R(s, \vec{a})$ for

taking the joint action \vec{a} in the state s . The initial state distribution is b^0 . The belief vector $b_i^t = [b_{i,1}^t \dots b_{i,N_S}^t]$ determines the belief of i -th agent about the state of the environment in time step t . In fact, b_i^t is a probability distribution over S such that $b_{i,n}^t$ specifies the belief of i -th agent that the state of the environment is s_n . The belief space is an N_S -dimensional space defined by the belief vector.

For infinite-horizon Dec-POMDP problems with the initial state distribution b^0 , the solution is a joint policy δ that maximizes the expected infinite-horizon discounted reward $E \left[\sum_{t=0}^{\infty} \gamma^t R(s^t, \vec{a}^t) | b^0 \right]$, where a discount factor γ ($0 \leq \gamma < 1$) limits the summation of rewards in the infinite-horizon.

Finding the optimal solution for the infinite-horizon Dec-POMDP may not be practical, because of unbounded number of steps [13]. Previous researches have tried to find a sub-optimal solution by using a bounded policy representation. The most common policy representation is finite state controllers (FSCs). Several approaches have presented to estimate the parameters of FSCs such as linear programming [7], nonlinear programming [8] and expectation-maximization [9], [10]. Value function is another approach to represent the policy in infinite-horizon Dec-POMDP problems [14]. In our previous work [12] we have introduced an incremental method to learn a fuzzy model as a value function. It generates a compact fuzzy rule-base as a solution that offers scalability for large MAS problems.

As stated before, obtaining minimal communication to coordinate the behavior of the agents is one of the main challenges in cooperative MAS problems. Therefore, several methods have been introduced to determine the communication strategy. Most of these algorithms work for finite-horizon Dec-POMDP cases [15], [6]. F. Wu et al. [1] introduced an online planning approach to reduce the computational complexity. To cope with limited bandwidth, the agents communicate only when history inconsistency is detected. The presented method in [16] calculates divergence between the agents' belief to evaluate communication. Since this method has considered an imprecise assumption for calculating belief divergence, it cannot accurately estimate the value of communication.

B. Incremental Learning

An incremental learning is a method that creates a model by recursively extracting required information from sequence of incoming data. This learning method is able to start learning "from scratch". Its parameters and structure are tuned incrementally according to current information without memorizing previous observation. Thus, the model can be created using low computational complexity and limited memory size. Evolving fuzzy [17] and neuro-fuzzy [18] systems are the most popular approaches for incremental learning. Shahparast et al. in [19] proposed two fast methods for adapting certainty factors of fuzzy rules, based on the reinforcement learning and reward and punishment. In [20] a simple and fast method is proposed that uses gradient decent to

tune the structure and parameters of a fuzzy classifier. D. Kangin et al. in [21] and [22] have introduced a group of incremental methods called TEDA that can be used for clustering, regression and classification. Incremental methods are also employed to find a policy for infinite-horizon Dec-POMDP. An incremental reinforcement learning algorithm is presented in [12] to create a compact fuzzy model as a solution of large MASs.

III. OUR PROPOSED METHOD

In this paper, we introduce a method to find a communication strategy for cooperative MAS problems in which the communication is expensive or limited. This method estimates the benefit of communication by computing the effect of communication on increasing accumulated reward for each situation. This can be used to obtain minimal communication that is necessary for successful joint behavior.

In this paper, we extend our previous method presented in [12]. In that method, each agent makes use of an individual fuzzy rule-base to interact with the environment. These rule-bases that map the belief space to the value of the actions, are created and tuned by an incremental reinforcement learning algorithm regarding experiences of the agents.

In this paper, two phases are considered, learning and execution phase. In the learning phase, communication between the agents is not limited and the algorithm freely shares the information of the agents to tune the communication strategy. However, there is limited bandwidth in the execution phase and the agents use the learned strategy to identify the situations where the communication can be beneficial to improve the performance.

A. Learning Phase

In this phase, the agents interact with the environment and in addition to tuning their behavior according to the response of the environment, the communication strategy is adjusted. To do this, each agent has an individual decision making system to select the best action in every time step and there is a shared *communication rule-base*, that is used to learn the benefit of communication for each situation.

The benefit of communication, Q_c^t , is computed by comparing the outcomes of two different agent-environment interactions in the particular state of the environment. Since the state of the environment is not available in Dec-POMDPs, we approximate it with the belief vectors of the agents. In each time step, once, the agents select the action without using communication and once again, the actions are selected after sharing information. The difference between the value of these two selected actions (i.e. immediate reward plus the expected accumulation of future rewards) determines the benefit of communication. Therefore, there is a tuple for each time step that contains two parts: particular situation of the environment, which specified by belief vector and Q_c^t , the benefit of communication for this situation. We call this tuple an *experience*. Fig. 2 illustrates the process of producing an *experience* in time step t .

Since the agents interact with the environment many times and an *experience* is achieved for each time step, there is a sequence of *experiences* (one element for each time step). The *communication rule-base* is created and tuned using this sequence. The sequence of *experiences* theoretically is infinite in infinite-horizon Dec-POMDP problems. Therefore, we have introduced an incremental algorithm to develop communication strategy. We describe the process of producing an *experience* and updating mechanism according to an *experience* in following two sub-sections.

1) *Producing an experience*: At each time step t , first, the agents interact with the environment, using only local information. Each agent updates its previous belief vector b_i^{t-1} to local belief vector b_i^t that is computed based on its previous action a_i^{t-1} and local observation o_i^t .

$$\forall s' \in S, b_i^t(s') = \frac{\sum_{\bar{a}_{xi}^{t-1} \in A_{xi}} \sum_{\bar{o}_{xi}^{t-1} \in O_{xi}} O(\bar{o}^t | s', \bar{a}^{t-1}) \sum_{s \in S} P(s' | s, \bar{a}^{t-1}) b_i^{t-1}(s)}{\sum_{s'' \in S} \sum_{\bar{a}_{xi}^{t-1} \in A_{xi}} \sum_{\bar{o}_{xi}^{t-1} \in O_{xi}} O(\bar{o}^t | s'', \bar{a}^{t-1}) \sum_{s \in S} P(s'' | s, \bar{a}^{t-1}) b_i^{t-1}(s)} \quad (1)$$

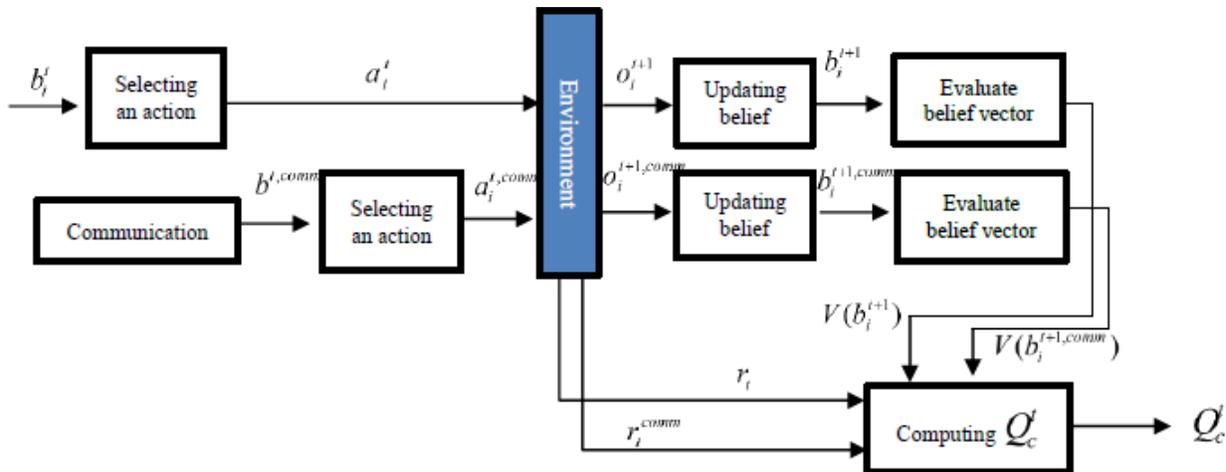


Fig. 2. The process of producing an *Experience* in time Step t .

Where, $\vec{o}_{\neq i}^t$ and $\vec{a}_{\neq i}^{t-1}$ are the joint observation and the joint action of all agents except agent i respectively. Also, $O_{\neq i}$ and $A_{\neq i}$ are all possible joint observations and all possible joint actions for the other agents, $\vec{o}^t = \langle o_i^t, \vec{o}_{\neq i}^t \rangle$ and $\vec{a}^{t-1} = \langle a_i^{t-1}, \vec{a}_{\neq i}^{t-1} \rangle$.

Then, according to b_i^t , the agent selects the best action. As stated before, we used our previous work presented in [12] to determine the behavior of the agents. In this method, each agent has individual fuzzy rule-base to estimate the value of the actions according to its belief vector. In fact, fuzzy rule-base of i -th agent determines $Q_i(b_i^t, a_m)$, the expected value of action m . At each time step, the agents estimate the value of their actions and perform the action having maximum value.

$$a_i^t = \arg \max_{a_m} Q_i(b_i^t, a_m) \quad (2)$$

After obtaining a_i^t , the same process is done to determine the appropriate joint action if the agents share their local information. To do this, the algorithm considers $b^{t,comm}$ as global belief vector and update it by using joint action \vec{a}^{t-1} and joint observation \vec{o}^t .

$$\forall s' \in S, b^{t,comm}(s') = \frac{O(\vec{o}^t | s', \vec{a}^{t-1}) \sum_{s \in S} P(s' | s, \vec{a}^{t-1}) b^{t-1,comm}(s)}{\sum_{s'' \in S} O(\vec{o}^t | s'', \vec{a}^{t-1}) \sum_{s \in S} P(s'' | s, \vec{a}^{t-1}) b^{t-1,comm}(s)} \quad (3)$$

Using global belief vector $b^{t,comm}$, each agent selects the best action $a_i^{t,comm}$:

$$a_i^{t,comm} = \arg \max_{a_m} Q_i(b^{t,comm}, a_m) \quad (4)$$

Therefore, there are two joint actions in each time step for interacting with the environment; if the agents communicate to each other, $\vec{a}^{t,comm}$ is selected and if they make decision based on the local information, \vec{a}^t is selected. The difference between the outputs of these two joint actions, determines the value of the communication in time step t .

Assume r_i^{comm} and $\vec{o}^{t+1,comm}$ are global reward and joint observation if the agents take joint action $\vec{a}^{t,comm}$; and if the agents perform joint action \vec{a}^t , they receive r_i and \vec{o}^{t+1} from the environment. The difference between the outputs of these two joint actions is calculated as follow:

$$Q_c^t = [r_i^{comm} + \gamma V(b_i^{t+1,comm})] - [r_i + \gamma V(b_i^{t+1})] \quad (5)$$

Where b_i^{t+1} is updated for each agent using a_i^t , o_i^{t+1} and (1), and also $b_i^{t+1,comm}$ is updated using $a_i^{t,comm}$, $o_i^{t+1,comm}$ and same equation. $V(b_i^{t+1,-})$ (i.e. $V(b_i^{t+1})$ or $V(b_i^{t+1,comm})$) is the estimated value of the incoming situation. In fact, $V(b_i^{t+1,-})$ estimates accumulated reward that will be achieved in the future steps. $V(b_i^{t+1,-})$ is easily obtained by one-step look-ahead:

$$V(b_i^{t+1,-}) = \max_{a' \in A_i} Q_i(b_i^{t+1,-}, a') \quad (6)$$

In this manner, whenever each agent has the same belief vector as b_i^t , the benefit of the communication is Q_c^t (i.e. communication can increase the accumulated reward by Q_c^t). Our proposed algorithm uses this tuple $[b_i^t, Q_c^t]$ as an *experience* to tune the *communication rule-base*.

2) *Updating mechanism*: In the learning phase, the agents interact with the environment many times and an *experience* is achieved for each time step. Hence, there is a sequence of *experiences* that our algorithm uses to create and tune the *communication rule-base*. The proposed method combines the information of *experiences* by clustering the similar *experiences*, in which center of each cluster identifies a communication rule. Since the number of *experiences* in the learning phase is huge, we introduce an incremental approach to cluster the *experiences*. In the following, we present the incremental process of tuning the communication strategy according to an *experience*:

Each rule specifies the benefit of communication for a region of belief space. The j -th rule in *communication rule-base*, R_j^{comm} , have a following form:

$$R_j^{comm} : \text{if } b_i^t \text{ is like } B_j^{comm} \text{ then } Q = Q_j^{comm} \quad (7)$$

Where $B_j^{comm} = [B_{j,1}^{comm} \dots B_{j,N_S}^{comm}]$ is a *reference belief vector* [12] of rule j that specifies the center of the region and Q represents the expected benefit of communication for this region.

Assume the i -th agent has an *experience* $[b_i^t, Q_c^t]$ in the time step t . The algorithm identifies the most similar reference belief vector to b_i^t . To do this, the similarity of b_i^t to the reference belief vector of all existing rules in the *communication rule-base* is computed as follows:

$$CosSim(b_i^t, B_j^{comm}) = \frac{\sum_{k=1}^{N_s} b_{i,k}^t B_{j,k}^{comm}}{\sqrt{\sum_{k=1}^{N_s} (b_{i,k}^t)^2} \sqrt{\sum_{k=1}^{N_s} (B_{j,k}^{comm})^2}} \quad (8)$$

Where $CosSim(b_i^t, B_j^{comm})$ is the cosine similarity of these two vectors.

If maximum similarity of b_i^t to the existing rules is less than threshold Sim_{min} , i.e. b_i^t considerably different with all reference belief vectors, so we consider $[b_i^t, Q_c^t]$ as a *new experience*. In this case, the proposed method adds a new rule to the *communication rule-base*, according to $[b_i^t, Q_c^t]$. The reference belief vector of the new rule is set to b_i^t ($B_{newRule}^{comm} \leftarrow b_i^t$) and the consequent part of the new rule is set to Q_c^t ($Q_{newRule} = Q_c^t$). It is noteworthy that if there is no rule in the *communication rule-base*, the same procedure is done to add the first rule.

Otherwise, if there is a similar reference belief vector to b_i^t , the nearest rule to b_i^t is determined:

$$w = \arg \max_j \{CosSim(b_i^t, B_j^{comm})\} \quad (9)$$

Where, w is the index of the most similar rule. Each rule is identified by averaging all similar *experiences* that agents have during the learning phase. Since the number of these *experiences* is huge, we use recursive formula to calculate the mean of group of similar *experiences*. For adjusting R_w^{comm} according to the *experience* $[b_i^t, Q_c^t]$, the antecedent of R_w^{comm} is updated regarding b_i^t by following recursive equation [21]:

$$B_{w(new)}^{comm} = \frac{(k-1)B_{w(old)}^{comm} + b_i^t}{k} \quad (10)$$

Where $B_{w(old)}^{comm}$ and $B_{w(new)}^{comm}$ are the reference belief vectors of R_w^{comm} , before and after updating, respectively. Similarly, the consequent of R_w^{comm} is updated as follow:

$$Q_{w(new)}^{comm} = \frac{(k-1)Q_{w(old)}^{comm} + Q_c^t}{k} \quad (11)$$

B. Execution Phase

The generated strategy is performed in the execution phase in which communication is limited. In this phase, the agents

estimate the benefit of communication and if it is recognized beneficial, the agents share their local information.

In each time step, the agents compute the benefit of communication according to its belief vector as follow:

Assume the belief vector of i -th agent in time step t is b_i^t . Firing strength of all rules in *communication rule-base* are calculated using:

$$\omega_j^t = \prod_{n=1}^{N_s} \mu_{B_{j,n}^{comm}}(b_{i,n}^t) \quad (12)$$

Where ω_j^t is the firing strength of the rule j . These firing strengths are then used to calculate the benefit of communication:

$$Q_{comm}(b_i^t) = \frac{\sum_{j=1}^{N_r} \omega_j^t Q_j^{comm}}{\sum_{j=1}^{N_r} \omega_j^t} \quad (13)$$

Where N_r is the number of rules in the *communication rule-base* and $Q_{comm}(b_i^t)$ denotes the benefit of communication from the perspective of i -th agent. This agent propagates communication request if the estimated benefit is more than predefined threshold C_{comm} :

$$Q_{comm}(b_i^t) > C_{comm} \quad (14)$$

The values of C_{comm} depends on the characteristics of each problem. In the real-world problems, this parameter can be set according to the percentage of access to the communication. Also, in an application with the communication cost, this parameter can be used to balance the communication costs with the coordination benefits.

If communication is available, each agent propagates its sequence of action-observation from previous communication, up to the current time step. By sharing this information, the belief vectors of all agents are equivalent and thus the coordinated behaviours are guaranteed. In the absence of communication, the agent postpones its request until the communication is allowed. By using this strategy, the behaviours of the agents maintain coordinated with little communication.

IV. EXPERIMENTAL RESULTS

We evaluated our proposed algorithm on several benchmark problems that have been widely used to rate multi-agent planning methods. These problems are Broadcast Channel [3], Meeting in a Grid 3x3 [4], Cooperative Box Pushing [5] and Stochastic Mars Rover [23]. We reported the accumulated discounted reward (Reward), percentage of communication (Comm. (%)) and the number of generated rules with different values of C_{comm} . In the real-world problems, C_{comm} can be set regarding the amount of access to the communication. Lower value of C_{comm} increases the

communication usage. In an application with the communication cost, C_{comm} can be used to balance communication costs with coordination benefits. The discount factor is set to 0.9 and the results are averages over 50 runs.

To the best of our knowledge, this is the first attempt to find the communication behaviour in infinite-horizon Dec-POMDP problems. Therefore, we compare the performance of our communication strategy to the full-communication (Full-Comm.) strategy as an upper bound and the no-communication (No-Comm.) strategy as a lower bound. Since in real-world MAS problems the communication is limited, the main purpose of the experiments is to test whether our proposed communication behaviour can help the agents to approach the performance of full-communication, while using little communication.

A. Broadcast Channel Problem

In the Broadcast Channel problem two agents are connected in a network. In each time step, only one of them can use the connection and sends its message. To avoid collision, each agent has to decide whether send a message or not. This problem has 4 states, 2 actions and 5 observations. The results in Table I show that Broadcast Channel problem is very simple such that the agents can easily cooperate. Therefore, the performance of the various percentage of communication is almost the same and different values of C_{comm} have no effect on the performance.

TABLE I. BROADCAST CHANNEL RESULTS

Broadcast channel	C_{comm}	Reward	Comm. (%)	No. of rules
$ S =4$	No-Comm.	9.1	0	-
$ A_i =2$	0.5	9.11	0.0	2
$ O_i =5$	0.1	9.18	83.16	1.16
	Full-Comm.	9.2	100	-

B. Meeting in a Grid Problem

In Meeting in a Grid problem, there are two agents on a 3×3 grid. They can move up, down, left or right, or stay on previous square. Each agent can sense whether there are walls around by noisy sensors with a 0.9 chance to perceiving the right observation. The goal of the agents is to spend as much time as possible on the same square. This problem has 81 states, 7 observations, 5 actions. The results in Table II show that low percentage of communication cannot significantly improve accumulated reward, however the performance of full-communication strategy can be achieved by making communication in almost half of time steps. Since the agents in Meeting in a Grid problem need the future planning information to cooperate, and in our method, the action-observation sequence is transferred, the proposed communication strategy cannot maintain the agents coordinated for a long time.

TABLE II. MEETING IN A 3×3 GRID RESULTS

Meeting in a 3×3 Grid	C_{comm}	Reward	Comm. (%)	No. of rules
$ S =81$	No-Comm.	4.19	0	-
$ A_i =5$	0.7	4.22	14.13	27.62
$ O_i =7$	0.5	5.71	57.99	27.94
	Full-Comm.	5.82	100	-

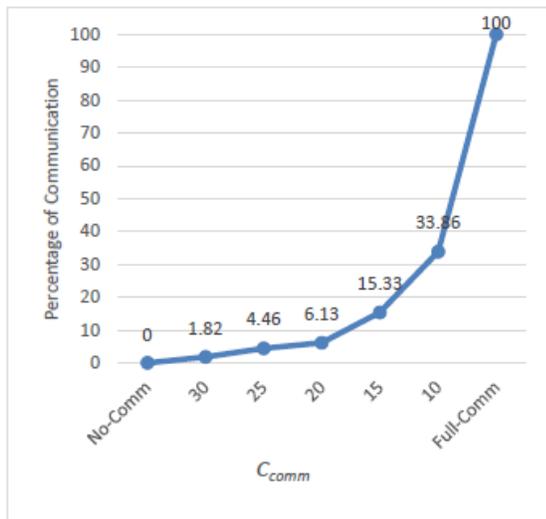
C. Cooperative Box Pushing Problem

In Cooperative Box Pushing problem, there are three boxes (two small and one large) on a 3×4 grid and two agents that can move the boxes. Each agent can push a small box alone. However, for moving the larger box, the agents need to cooperate. Whenever one of the boxes reaches into a goal area, a trial ends. If it is one of the small boxes, the agents gain a reward of +10, and if the large box move into the goal area, they get a reward of +100. However, if a box smashes into a wall or the large box is pushed by one agent, a penalty of -5 is received. The Box Pushing problem has 4 actions, 5 observations, 4 goal states and 96 non-goal states (100 states in total). According to the definition of this problem, communication has a significant impact on the performance. The reported results in Table III show the proposed communication strategy did significantly improve the performance with low percentage of communication. While the achieved accumulated reward with no communication is 177.11, this value can be increased to 218.97 by communicating in only 6.13% of time steps. Also, the accumulated reward has reached 225.19 by communicating in one third of time steps whereas it is 232.25 for the full-communication case.

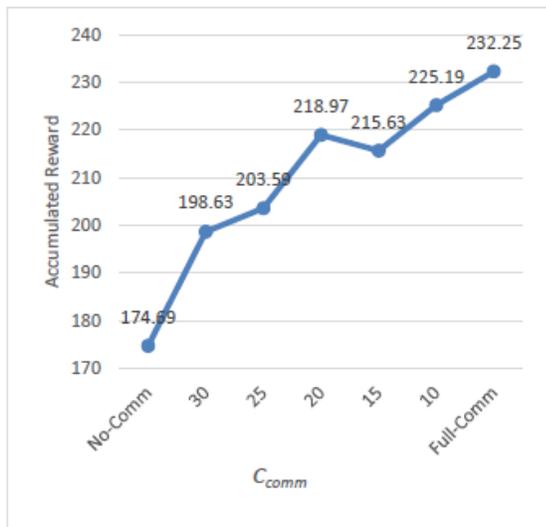
Fig. 3 demonstrates the effect of different values of parameter C_{comm} on the percentage of communication and the accumulated reward in solving Cooperative box pushing problem. In order to better illustration of the performance of our method, the values of the accumulated reward are shown between the achieved reward of the No-Comm. strategy as a lower bound and the Full-Comm. strategy as an upper bound. As stated before, the percentage of communication and accumulated reward are increased by decreasing C_{comm} . Moreover, regarding these figures it is obvious that the accumulated reward is significantly increased with a small increase in percentage of communication.

TABLE III. COOPERATIVE BOX PUSHING RESULTS

Cooperative box pushing	C_{comm}	Reward	Comm. (%)	No. of rules
$ S =100$	No-Comm.	177.11	0	-
$ A_i =4$	30	198.63	1.82	26.34
$ O_i =5$	20	218.97	6.13	26.34
	10	225.19	33.86	26.56
	Full-Comm.	232.25	100	-



(a)



(b)

Fig. 3. The Effect of C_{comm} on (a) the Percentage of Communication and (b) the Accumulated Reward in Cooperative Box Pushing Problem.

D. Mars Rover Problem

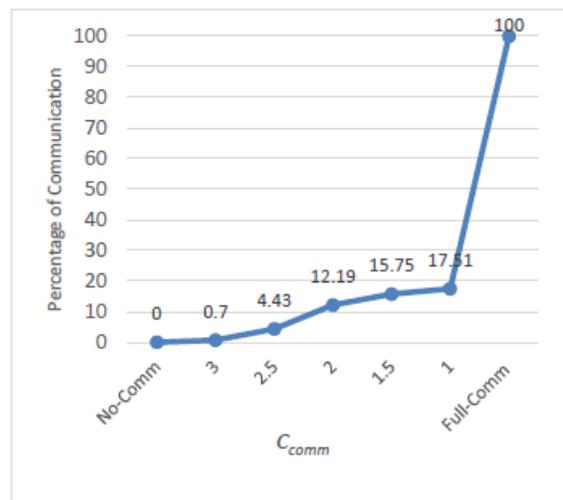
We evaluate the performance of our proposed method with a larger problem, Mars Rover problem. In This problem, there are two rovers experimenting at a 2×2 grid by independently drilling or sampling at each site or moving around. Two of the sites just need one agent to sample, while in the other sites, both agents must drill at the same time in order to get the maximum reward. The agents get a large penalty, if a site is drilled while it only needs to be sampled. When at least one experiment is performed at each site, the problem is reset. This problem has 256 states, 6 actions and 8 observations. As can be seen from Table IV, proposed communication strategy did very well for Mars Rover problem as a large MAS problem. The method achieves almost the same performance as the case of full-communication by making communication in less than one fifth of time steps (17.51%).

We have also demonstrated the results of accumulated rewards and the percentage of communication with different

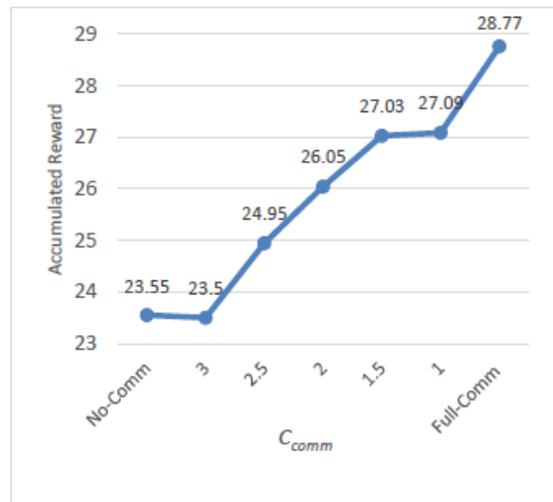
values of C_{comm} in solving Mars rover problem in Fig. 4. Fig. 4(a) illustrates the effect of C_{comm} on the percentage of communication and Fig. 4(b) shows the effect of this parameter on the accumulated reward. Again, in Fig. 4(b), the values of accumulated reward are shown between the reward of the No-Comm. strategy and the Full-Comm. strategy as the lower and upper bound, respectively. Fig. 4 clearly shows that with a small increase in percentage of communication, the accumulated reward is significantly increased.

TABLE IV. MARS ROVER RESULTS

Mars Rover	C_{comm}	Reward	Comm. (%)	No. of rules
S =256 A _i =6 O _i =8	No-Comm.	23.55	0	-
	3	23.5	0.7	8.06
	2	26.05	12.19	8.02
	1	27.09	17.51	8.24
	Full-Comm.	28.77	100	-



(a)



(b)

Fig. 4. The Effect of C_{comm} on (a) the Percentage of Communication and (b) the Accumulated Reward in Mars Rover Problem.

To summarize, our proposed algorithm to develop the communication strategy, performed very well in all the benchmark problems. Using this strategy can heavily reduce the amount of communication necessary for successful coordinated behaviour.

V. CONCLUSION

We introduced an algorithm to develop a communication strategy for cooperative multi-agent systems in which the communication is limited. This strategy identifies best situations for making communication in MASs modelled by infinite-horizon Dec-POMDP. This communication policy is developed centralized in a training phase, which the communication is not restricted. The agents use this policy decentralized in a test environment that the communication channel is limited. Our method generates a fuzzy model to approximate the benefit of communication for each situation. The agents can use this fuzzy model to obtain minimal communication that is necessary for coordinated behavior. We also introduced an incremental method to create and tune this fuzzy model. Our incremental method has reduced the high computational complexity of the multi-agent systems by constructing a compact fuzzy rule-base. We used several standard benchmark problems to evaluate the performance of our proposed method. Experimental results show that this communication strategy can help the agents to achieve almost the same performance as the full-communication strategy by using little communication. Therefore, in the real-world MAS problems that the communication is usually limited, our proposed algorithm can heavily reduce the amount of communication necessary for successful coordinated behaviour.

Many AI domains can take advantage of MAS design such as multiple mobile robots and disaster response teams. Developing a group of intelligent players or agents in video games is another interesting field in AI research. In our future work, we intend to customize our incremental model to create human-like players for real-time strategy games who can act and react intelligently against virtual environment and even real players.

REFERENCES

- [1] E Wu, S. Zilberstein and X. Chen, "Online Planning for Multi-Agent Systems with Bounded Communication," *Artificial Intelligence*, vol. 175, no. 2, p. 487–511, 2011.
- [2] D. S. Bernstein, R. Givan, N. Immerman and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," in *Mathematics of Operations Research* 27, 2002.
- [3] D. S. Bernstein, E. A. Hansen and S. Zilberstein, "Bounded policy iteration for decentralized POMDPs," in *Proceedings of the 19th international joint conference on Artificial intelligence*, 2005.
- [4] C. Amato, J. S. Dibangoye and S. Zilberstein, "Incremental Policy Generation for Finite-Horizon DEC-POMDPs," in *Proceedings of the 19th International Conference on Automated Planning and Scheduling*, Thessaloniki, Greece, 2009.
- [5] S. Seuken and S. Zilberstein, "Improved Memory-Bounded Dynamic Programming for Decentralized POMDPs," in *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, Vancouver, British Columbia, 2007.
- [6] M. Roth, R. Simmons and M. Veloso, "Reasoning about joint beliefs for execution-time communication decisions," in *AAMAS '05 Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, 2005.
- [7] D. S. Bernstein, C. Amato, E. A. Hansen and S. Zilberstein, "Policy Iteration for Decentralized Control of Markov Decision Processes," *Journal of AI Research (JAIR)*, vol. 34, pp. 89-132, 2009.
- [8] C. Amato, D. S. Bernstein, and S. Zilberstein, "Optimizing fixed-size stochastic controllers for POMDPs and decentralized POMDPs," *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)*, vol. 21, no. 3, p. 293–320, 2010.
- [9] J. K. Pajarinen and J. Peltonen, "Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning," in the *25th Annual Conference on Neural Information Processing Systems (NIPS 2011)*, 2011.
- [10] A. Kumar and S. Zilberstein, "Anytime Planning for Decentralized POMDPs using Expectation Maximization," in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, California, 2010.
- [11] R. Sharma and M. T. J. Spaan, "Bayesian-Game-Based Fuzzy Reinforcement Learning Control for Decentralized POMDPs," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 4, pp. 309 - 328, 2012.
- [12] S. Hamzeloo and M. Zolghadri Jahromi, "An incremental fuzzy controller for large dec-POMDPs," in *Artificial Intelligence and Signal Processing Conference (AISP)*, Shiraz, Iran, 2017.
- [13] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*, Springer International Publishing, 2016.
- [14] H. Kurniawati, D. Hsu and W. S. Lee, "Efficient point-based POMDP planning by approximating optimally reachable belief spaces," in *In Proc. Robotics: Science and Systems*, 2008.
- [15] R. Emery-Montemerlo, *Game-theoretic control for robot teams*, Doctoral Dissertation, Robotics Institute, Carnegie Mellon University, August 2005.
- [16] S. A. Williamson, E. H. Gerding and N. R. Jennings, "Reward shaping for valuing communications during multi-agent coordination," in *AAMAS '09 Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, Budapest, Hungary, May 10 - 15, 2009.
- [17] P. P. Angelov and X. Zhou, "Evolving Fuzzy-Rule-Based Classifiers From Data Streams," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 6, 2008.
- [18] S. Schliebs and N. Kasabov, "Evolving spiking neural network—a survey," *Evolving Systems*, vol. 4, no. 2, p. 7–98, 2013.
- [19] H. Shahparast, S. Hamzeloo and M. Zolghadri Jahromi, "A Self-Tuning Fuzzy Rule-Based Classifier for Data Streams," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 22, no. 2, 2014.
- [20] H. Shahparast and E. G. Mansoori, "An online fuzzy model for classification of data streams with drift," in *Artificial Intelligence and Signal Processing Conference (AISP)*, Shiraz, Iran, 25-27 Oct. 2017.
- [21] D. Kangin, P. Angelov and J. A. Iglesias, "Autonomously evolving classifier TEDAClass," *Information Sciences*, vol. 366, p. 1–11, 2016.
- [22] D. Kangin and P. Angelov, "Evolving clustering, classification and regression with TEDA," in *International Joint Conference on Neural Networks (IJCNN)*, 2015.
- [23] C. Amato and S. Zilberstein, "Achieving goals in decentralized POMDPs," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, 2009.

OpenSimulator based Multi-User Virtual World: A Framework for the Creation of Distant and Virtual Practical Activities

MOURDI Youssef, SADGAL Mohamed, BERRADA FATHI Wafaa, EL KABTANE Hamada

Faculty Semlalia
University Cadi Ayyad
Marrakesh, Morocco

Abstract—The exponential growth of technology has contributed to the positive revolution of distance learning. E-learning is becoming increasingly used in the transfer of knowledge where instructors can model and script their courses in several formats such as files, videos and quizzes. In order to complete their courses, practical activities are very important. Several instructors have joined Multi-User Virtual World (MUVW) communities such as SecondLife, as they offer a degree of interrelated realism and interaction between users. The modeling and scenarization of practical activities in the MUVWs remains a very difficult task considering the technologies used by these MUVWs and the necessary prerequisites. In this paper, we propose a framework for the OpenSimulator MUVWs that can simplify the scenarization of practical activities using the OpenSpace3D software and without requiring designers to have expertise in programming or coding.

Keywords—E-Learning; multi-user virtual world; practical activities; OpenSimulator; virtual reality; virtual laboratories

I. INTRODUCTION

In the current circumstances, the use of Information and Communication Technologies (ICT) in both 'distance' and face to face training and particularly in the form of virtual learning environment such as learning management systems (LMS), has become a necessity for universities, high schools and even training companies, consequently they tend to virtualize their education and training system [1], this is due to all the features offered by these virtual environments including document sharing, collaboration, discussion, and communication between learners and their teachers [2].

It should be noted that providing a quality educational experience for all courses type is not easy, some of them like engineering courses require practical activities (PA) and therefore the laboratory concept is very important in the learning process [3][4].

In fact, PAs enable learners to develop useful skills for their professional career, to learn how to handle and use things accurately, to discover new things, and even to test the veracity of an idea. For teachers, practical activities are the best way to build knowledge, to understand facts and scientific theories and to stimulate learner's motivation. This can be done by performing concrete experiences, developing psychomotor

skills and/or illustrating scientific approaches to solve problems.

A huge quantity of solutions has been attempted in order to propose approaches for teachers to integrate PAs in distance learning. The first was the remote-control laboratory [5][6][7][8], then the use of videos and finally virtual laboratories [9]. However, the latest proposal is very superficial, and very limited.

At this level, several challenges arise following the providing of PAs remotely and which come in two main issues: the first is to have a scalable and open environment for tutors to design, model practical activities and set it accessible to learners. The second challenge is to have a collaboration and communication environment, with a great degree of realism and to give the students the opportunity to take their practical activities by immersing them as much as possible in a learning situation. These requirements have pushed researchers in education domain to orient themselves towards the 3D environments as multi-user virtual worlds (MUVW) like Second Life, OpenSimulator and others, in order to measure their efficiencies and their capacities to meet the needs of teachers and learners [10][11][12][13][14]. Therefore, a set of 3D simulation solutions in MUVE have emerged like in [15][16]. Even if this approach offers very immersive environments with a high degree of realism, its implementation remains difficult to achieve. As a teacher, this difficulty occurs on four levels, namely to model, animate, interact and share with learners a 3D scene containing objects.

This paper presents a solution for developing practical activities based on OpenSimulator MUVWs, specifically using the open source software OpenSpace 3D [17]. The purpose is to create virtual reality applications by using a simple graphical interface and minimizing the code as much as possible. The proposed solution is a bridge between the 3D scenes created with openSpace 3D and the worlds of OpenSimulator [18] which aims to make the creation of virtual practical activities easier for teachers.

This paper is divided into several sections and is organized as follows: Section II presents the various works that have been done to ensure PAs in e-learning and illustrates a comparative study of the different proposed approaches. In Section III, we present the architecture of the proposed framework as well as it

sub-modules, and particularly the transformation module which is fully described in Section IV. Section V presents a case study to test the proposed framework and thus it shows the results obtained before a discussion and a conclusion section (Section VI).

II. LITERATURE REVIEW AND MOTIVATION

Distance education has seen a great evolution as Mari Lahti and her colleagues show in [19]. This development is due to the true potential of the e-learning concept. In fact, distance learning is a very useful training tool for researchers who have problems related to their geographic location or adult learners who have work commitments, family constraints or other obstacles [19]. Contrariwise, we may be confronted with topics that require PAs, especially the courses that refer to the fields of engineering and science [20]. Moreover, experiments can emphasize prominent information or remove confusing details [20].

Several approaches have come to solve the problem of remote practical activities. In this section, we present these solutions and we try to show and discuss their weaknesses.

A. Remote Control Laboratory (RCL)

The principle is to enable different learners to remotely control real equipment using a software interface as in [21][22][23][24]. During the experiment, the results obtained are retrieved and stored in a database, or displayed on the interface that the learner uses to communicate with the remote laboratory.

Fig. 1 illustrates a basic model of RCL based approaches. In remote labs, students manipulate physical apparatus and get real data from physical experiments [20]. This solution is very limited for several reasons:

- The investment to ensure the purchase and the maintenance of the real hardware is always a problem in such cases. On the other hand, if a device fails during testing, the student is obliged to contact the person responsible for maintenance. The presence of a technician is mandatory, and the resumption of the workshop can take hours.
- If this solution offers us the ability to control robots or remote machines, it cannot ensure practical workshops

(for example, in chemistry). It can only be used in teaching simple engineering experiments.

- The most prevalent downside is the large number of experiments in relation to practical activities that can run simultaneously. The major problem in education, especially at the university, is the massive number of enrolled students, and therefore the number of required workshops and investment they need.
- This can cause feelings of isolation for the student and hence reduces his/her motivation [21].

B. The use of Videos

To best fit the needs of this problematic and the different limits of the solution for remote control of the real hardware, another proposal has emerged. The idea is to record one version of the practical activity and share it with the learner when it reaches the stage of practice in his learning course [25].

This solution solves, actually, a few of the previous limits, but it puts the learner in a passive learning state, due to the lack of interactivity, and especially for learners who need to be immersed in the learning experience in order to understand the different notions.

C. Virtual Laboratories and Simulators

Currently, either for 'distance' or 'face-to-face' education, the use of virtual laboratories (VL) experienced a great growth in various disciplines, namely the modeling and simulation of complex systems in physics [26][27], chemistry [11][28], biology [29] [30][31] and even robotics and engineering [32] [33].

This growth is due to several reasons. Indeed, the use of such materials has very significant advantages. In [34], the authors show that VL has several objectives on different levels such as instrumentation, models, equipment, data analysis, design, creativity, psychomotor, security, communication and teamwork.

Nevertheless, these environments are curtailed. A virtual laboratory is a set of virtual objects modeling real objects. Therefore, their extension is very difficult. If the tutor needs to add an object for some reason, he/she must wait for the release of the next version, and there can be no assurance that these processes will fit the expectations of the users.

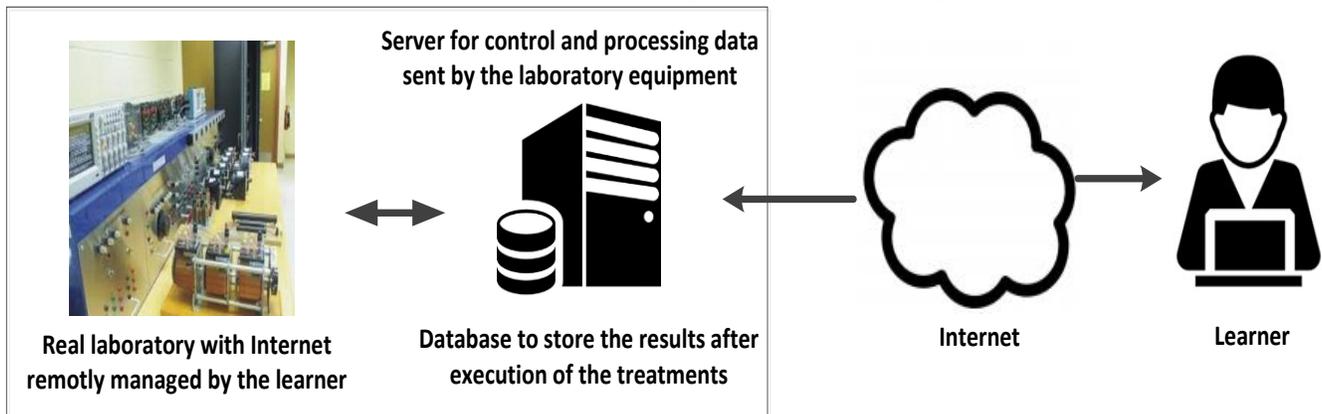


Fig. 1. Remote Control of Real Equipment Model.

D. The Use of Multi-User Virtual Worlds (MUVW)

Another way to ensure simulations, which becomes increasingly adopted [35], is the use of MUVWs. This growth is evidenced by the significant amount of research on this topic and the number of affected areas.

First, a multi-user virtual world, defined in [35], is an online environment, allowing multiple users simultaneous access to virtual resources and contexts where each user is represented by an avatar. Users can interact, communicate and share experiences that may be similar to the real world. Currently there is a large number of MUVW platforms such as Second Life, Open Simulator, Multiverse, and many others for rotating MUVWs.

Today, the features listed in the definition are no longer restricted to MUVWs, but they are redirected to be used in other areas, namely science [36], medicine, engineering [37] and many other disciplines [38].

E. Comparative Study of Existing Approaches

To offer a solution that meets as much as possible the needs of users, a comparison of existing solutions is required. The following table (Table I) illustrates the main differences between them, based on criteria that we considerate pertinent.

It can be seen that VLs share with MUVWs the same strengths in terms of the realization and the learner pedagogical situation. For against, the use of MUVWs leave a great level of immersion as well as the teachers and learners interaction, especially when it comes to 3D virtual worlds.

VLs remain limited to the objectives for which they were developed. Therefore, if the teacher wants to modify the proposed practical activity, s/he is compelled to code (knowing that all the proposed solutions are not open source), which requires the expertise of experts in this kind of programming cases.

TABLE I. COMPARATIVE STUDY OF EXISTING APPROACHES

<i>Solution</i>	<i>RCL</i>	<i>Videos</i>	<i>VL</i>	<i>MUVW</i>
<i>Criterion</i>				
<i>Realization</i>	Very expensive	Less expensive	Less expensive	Less expensive
<i>Maintenance</i>	Very expensive	No cost	--	--
<i>Level of immersion</i>	Unavailable	Unavailable	Limited	Available
<i>Evolution</i>	--	--	Not always possible	Possible
<i>Cost</i>	Very expensive	--	Expensive	Less expensive
<i>Learner pedagogical situation</i>	Liability	Liability	Asset	Asset
<i>Interaction</i>	Unavailable	Unavailable	Limited	Excellent

III. THE PROPOSED APPROACH

A. Presentation

The primary focus in this work is to design and implement a computer system addressing the issue of simulation by providing an approach for the preparation of virtual PAs and meeting the limits of existing approaches. At this level, several questions arise:

- How can the teacher create practical activities while having an extensible environment?
- How to create interactive objects, upload them and send them to students, knowing that not all teachers are computer scientists or either have programming skills?
- Is there a way to realize solutions and contextual factors in the different workshops between students?

In this section, we present our project: a simulation system composed of several modules that we consider important to model, animate and realize a practical activity. This section is divided into several parts, the first describes the overall architecture of the proposed system, and the second presents the approach used to create and realize a practical activity, while the third part explains the architecture's modules and the technologies used. Finally, the fourth part details an important module of the solution, which is the transformation module and represents the proposed system base.

B. Global Architecture

As we have mentioned, the framework we propose is divided into several modules, as shown in Fig. 2. The first module is a 3D object design tool. The second is a scripting tool for animating and making the interaction with the objects provided by the graphic designer. The transformation module manages the conversion of the language and formats generated after the teacher animates the objects of PAs to the language and format supported by the MUVW, this module is practically the cornerstone of this project.

The Fig. 2 shows also the actors involved in the system in its entirety. On one hand, there is the designer who, with his/her expertise, helps to design the PA. In a second hand, there is the teacher, as an essential actor in a teaching situation. The system must provide him/her with a set of features allowing him/her to prepare, monitor and control PAs. Eventually, the student is the targeted actor and the most important one. The system must offer him/her a sophisticated environment facilitating the communication with the other students, in addition to his/her interaction with the workshop and the objects. Thus, it will help assimilate easily the different concepts applied in a workshop.

C. The System Modules and Technology

1) *3D modeling tools*: The main function of this module is to model 3D objects to animate them later in OpenSpace3D. In fact, a specific tool does not limit us, any 3D modelling software can do the trick, we quote Blender, Maya, 3D Max, etc.

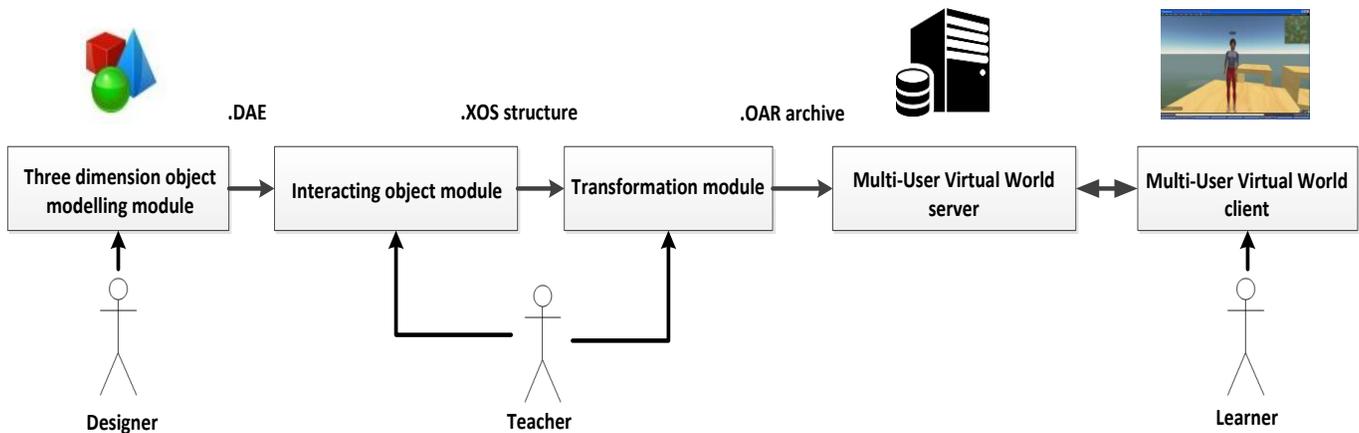


Fig. 2. Global Architecture.

2) *Scripting environment of practical activities*: It is the main tool with which the teacher, with the help of a graphic designer, can prepare practical workshops, and animate events (e.g., click, and double click). This module is the OpenSpace 3D software, which is an open source platform for developing Virtual and Augmented Reality projects. It allows full build of interactive 3D scenes graphically rich without entering any line of code. On the other hand, it also has many functions ready for use. Achieving an OpenSpace3D implementation therefore is integrating different functions and defining their reciprocal interactions.

OpenSpace 3D is based on plugIt principle to animate 3D objects. PlugIt is a function package already developed with Scol language [39] and a very simple way to animate 3D objects. The tool offers a sophisticated interface, divided into 3 parts (Fig. 3): the first one lists the imported 3D objects in the project, the second part shows the object's scene in 3D and the last part is where the user animates the objects with PlugIts just by using links between the event and the action. A link is a connection between an event function and an action of another or the same function. After the animation of the scene, the user saves his/her project under XOS format or s/he can import it towards a Scol project or a standalone application.

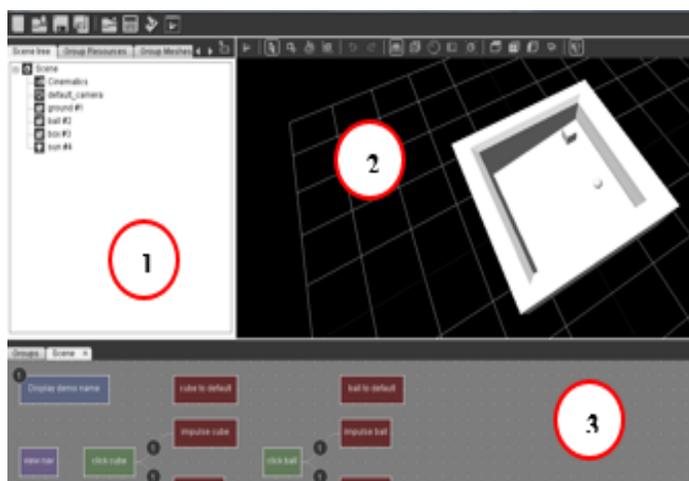


Fig. 3. OpenSpace 3D Environment.

OpenSpace 3D is not only limited to the creation of virtual reality applications, but also offers a set of Plugit dedicated to the management of augmented and mixed reality. It also offers the possibility of use and integration of several sensors, for example Leap Motion and Kinect Xbox. It can facilitate the development of applications and games which can be oriented to the educational area.

3) *MUVW Module*: This module will, to some extent, play the role of virtual laboratory. It represents the environment where the learner will make his/her PAs and communicate with other users whether they are students or teachers. The choice was made on the OpenSimulator platform[40], an open source project under the BSD license. It aims to develop a functional platform for virtual worlds capable of supporting multiple clients and servers, all in a heterogeneous grid structure. It allows great freedom concerning the development of specific applications, hosting and controlling overall costs for economic and efficient projects. This is partly justified, by the free license contrary to proprietary solutions.

Furthermore the interface (3D client browsers) already incorporates comprehensive tools for creating content accessible to neophytes (3D objects from simple primitive forms, scripts for programming, management of the 3D environment, etc.). Furthermore, design professionals who are familiar with softwares such as Blender, Maya, and 3DS Max, etc., can also incorporate 3D content created from these professional softwares. Finally, some collaboration tools are available in Open Simulator like Vivox, the leader in voice conference, Shared Media that allows a simple integration of multiple, rich and diverse media (web pages, video, images, etc.). These tools bring Open Simulator to an unparalleled level of functionality compared to other solutions.

4) *Transformation module*: This module is the processing interface that adapts 3D objects animated by the practical workshops' preparation module (OpenSpace 3D) to the OAR format used by Open Simulator. This transformation module, fully explained in the next section, exploits the transformation rules and the code generation to LSL (Linden Scripting language), the animated language adopted by Open Simulator to animate 3D objects.

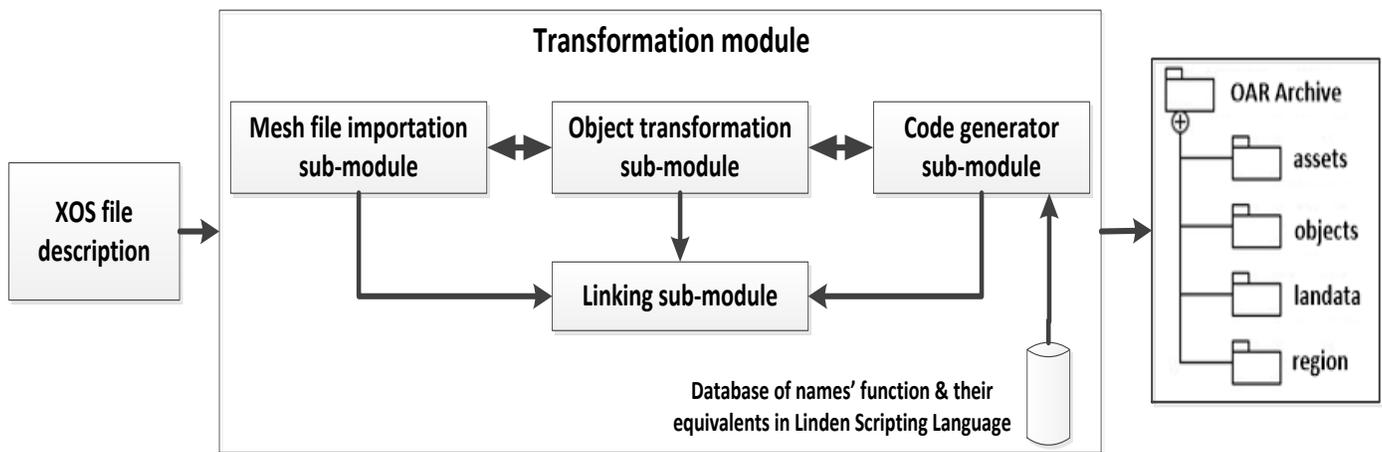


Fig. 4. Transformation Module Architecture.

D. Transformation Module Details

1) *Architecture*: For each scene of objects, OpenSpace3D generates an XML file with XOS extension that contains a description of the entire scene, events, and the location of 3D objects on the hard disk. The import unit searches the XML list files of 3D objects used in the practical workshop, and then it loads the objects in the OpenSimulator objects database. The transformation engine provides the correspondence between the activities defined in the XML file and generates the relevant objects and LSL Script based on defined transformation rules, and at the end, it loads the scripts on the server. The Fig. 4 illustrates the transformation mechanism from the XOS format into the OAR archive format.

We have segmented the transformation of the XOS files into the OAR format in several parts. The first step is to import the necessary resources for the construction of 3D objects and thus of the scene. The second step is searching for the

positioning information of each object on the stage and its characteristics. The third step is to transform the animations by searching in the database for the functions in LSL equivalent to those mentioned in the XOS file. The last step is to ensure the link between the objects, their characteristics and the animations (interactions).

2) *Meta-model of XOS*: After animating 3D objects of practical activity in OpenSpace 3D, the teacher saves the entire scene as a single compressed XML file with the XOS extension. The main role is to describe all objects in the scene, cameras, reference objects and their positions in the scene on one side, and the different events and own animation of each object in the other side. The Fig. 5 presents a part from an example of the XOS file.

We are based on several examples developed under Open Space 3D to propose a general meta-model of the XOS description. The Fig. 6 shows a simplified Meta model of an XOS file.

```
<project author="OpenSpace+3D+Objects+group+based+on+OgreMax+Scene+Exporter+@28vww@2eogremax@2ecom@29" formatVersion="1@2e22@2e0">
  <scene py="1" px="3">
    <resources>
      <resource path="assets@2fmodels@2fconverted@2fchute@5flibre@5fcollada@2fmaterials@2fchute@5flibre@5fcollada@2ematerial" type="material"/>
      <resource path="assets@2fmodels@2fconverted@2fchute@5flibre@5fcollada@2fmeshes@2fchute@5flibre@5fcollada@5ftext@5f0042@2emesh" type="mesh"/>
      ...
    <plugins>
      <plugin name="object+click" source="objects@2fobjectclick@2fobjectclick@2exml">
        <instance py="1" px="9" name="click+on+the+0@2e2+value" comment="">
          <param name="object"><![CDATA[14.chute_libre_collada_Text19]]></param>
          <param name="material"><![CDATA[chute_libre_collada_Material_0010]]></param>
          <param name="enablemat"><![CDATA[1]]></param>
          <param name="cursor"><![CDATA[1]]></param>
          ...
        </instance>
      </plugin>
    </plugins>
  </scene>
</project>
<mesh visibilityFlags="0xFFFFFF" nbLODlevel="0" renderQueue="50" renderingDistance="" indexMaterials="false" receiveShadows="false" castShadows=
  <subentities>
    <subentity defaultMaterial="chute@5flibre@5fcollada@5fMaterial@5f0010" materialName="chute@5flibre@5fcollada@5fMaterial@5f0010" index="0"/>
  </subentities>
  <scale z="3@2e089895" y="0@2e484955" x="0@2e484955"/>
  <position z="3@2e104373" y="0@2e282935" x="@2d2@2e195202"/>
  <rotation qw="1@2e0" qz="0@2e0" qy="0@2e0" qx="0@2e0"/>
```

Fig. 5. Example of an XOS File.

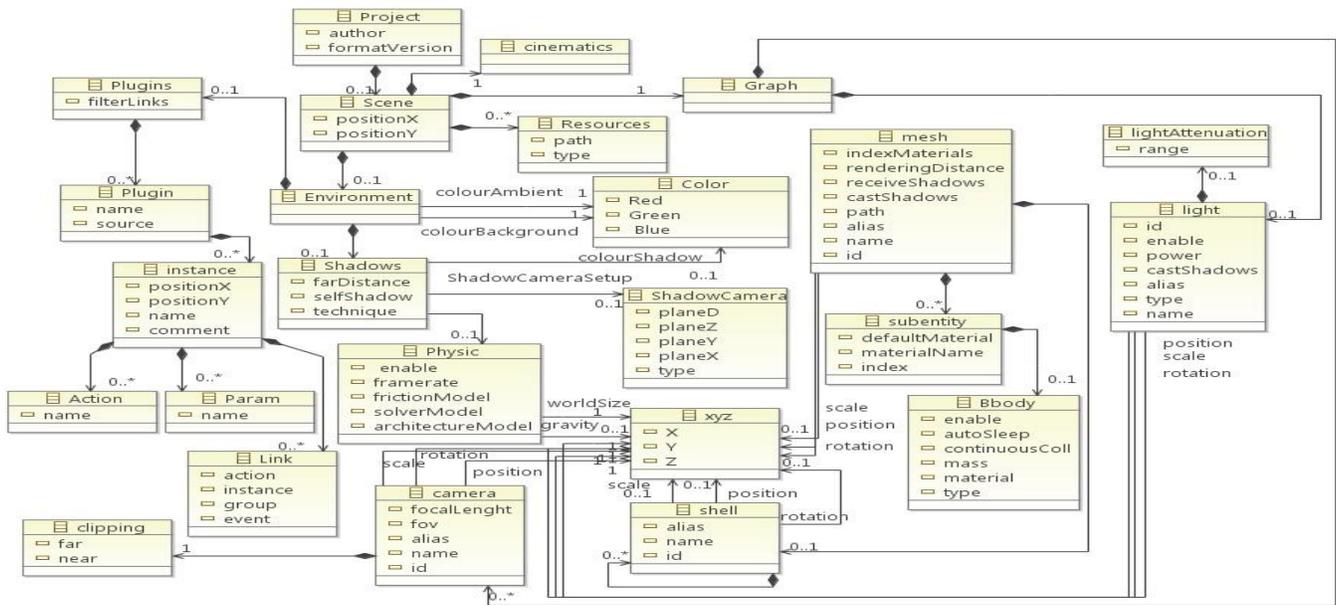


Fig. 6. Simplified Meta-Model of XOS.

3) Meta-model of OAR format

a) *The OAR model:* The OpenSimulator Archive (OAR) function as it is introduced in the official web site of OpenSimulator has existed since OpenSimulator 0.5.9. The facility does a similar job to load-xml2/save-xml2 and it saves prims so that they can be reloaded later.

However, OpenSimulator (OAR) archives go further in this direction, since they can back up all the inventory data needed to fully restore terrain, region parcel data, object textures and their inventories when they are loaded on a completely different system using a different asset database.

An OAR file is a “gzipped” tar file in the original Unix tar format. This can be extracted and created with standard tools. The Fig. 7 shows the OAR structure and components.

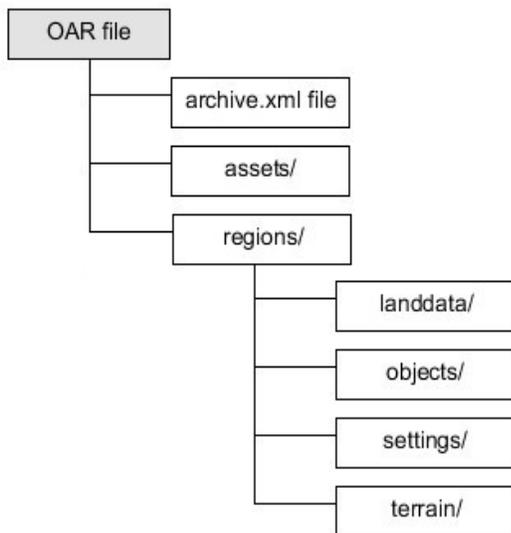


Fig. 7. OAR Format.

As we can see in Fig. 7, the OAR format is a set of directories; each one has a specific role. In our approach, OAR is the exchange format between the teacher and different learners. As following, a global description of the OAR format:

- **Archive.xml:** This is the archive control file. It contains a major and minor version number, to allow compatibility with future format changes.
- **Assets/:** This directory contains all the assets in the archive. The assets for all the regions are stored in the same directory because assets are often shared. Each filename has the following format <uuid>_<asset type>.<asset extension>. The uuid section must always be present and form a valid uuid - it is used directly for that asset. The asset type and asset extension are used to identify the type of asset and the asset extension allows the asset to be associated with different editors on platforms such as Windows. For instance, a script will always have the asset type and extension script.lsl. A full list of asset types and extensions can be found in the file.
- **Landdata/ :** This directory contains all the parcels in the region. Information for each parcel is stored in a separate file in XML format.
- **Objects/ :** Each individual file is an object in the region (where an object [linkset] can be composed of many prims). The file format used is OpenSim's XML2 format. Each filename has the following structure by default.
- **Settings/ :** This contains the region settings information for the region in XML format. The filename will be the same as the region name.
- **Terrains/ :** This contains the terrain file for the region, stored in RAW format. The filename must end with .r32.

```

<SceneObjectGroup>
  <SceneObjectPart xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
    <AllowedDrop>false</AllowedDrop>
    :
    <GroupPosition>
      <X>126.343</X>
      <Y>125</Y>
      <Z>29.08239</Z>
    </GroupPosition>
    <OffsetPosition>
      <X>0</X>
      <Y>0</Y>
      <Z>0</Z>
    </OffsetPosition>
    :
    <Color>
      <R>0</R>
      <G>0</G>
      <B>0</B>
      <A>255</A>
    </Color>
    :
    <Shape>
      <ProfileCurve>5</ProfileCurve>
      <TextureEntry>1VvnrTLQ+2SC0fK7RVGXwAAAAAAAAAIEEAAAAGQAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA</TextureEntry>
      <ExtraParams>AA==</ExtraParams>
      <PathBegin>0</PathBegin>
      <PathCurve>32</PathCurve>
      <PathEnd>0</PathEnd>
    </Shape>
  </SceneObjectPart>
</SceneObjectGroup>
  
```

Fig. 8. Example of Object Presentation in OpenSimulator using XML Model.

To identify equivalences between XOS model and the 3D objects OpenSimulator, we have also created a meta-model.

b) *Object's model on OpenSimulator*: The most important in this first stage is to export objects with animation script from XOS format to OAR archive. Therefore, we focused on the objects directory to model 3D object and on the assets directory to create script animation and affect each one to its 3D object. Firstly, we developed a meta-model of how 3D objects are modelled on OAR format. Each 3D object is defined with a file in the objects directory.

The Fig. 8 shows a part from an example of object presentation in OpenSimulator using XML model.

We are based on several examples developed under Open Space 3D to propose a general meta-model of the XML 3D

object description in OpenSimulator. The Fig. 9 shows the schematic simplified representation of a 3D object in an OpenSimulator's scene.

Each object is a "SceneObjectGroup" in the representation's file, this SceneObjectGroup is a set of SceneObjectPart, each SceneObjectPart has <x,y,z> coordinates. If SceneObjectGroup is a set of many mesh objects, a SceneObjectPart is defined in this case as a set of SceneObjectPart which is, in turn, defined in an OtherPart tag.

c) *Linden Scripting Language (LSL)*: Is the programming language used by residents of OpenSim. LSL has a syntax similar to C language and allows objects to control the behavior of in-world objects of Second Life from the Internet via email, XML-RPC, and most recently, HTTP requests.

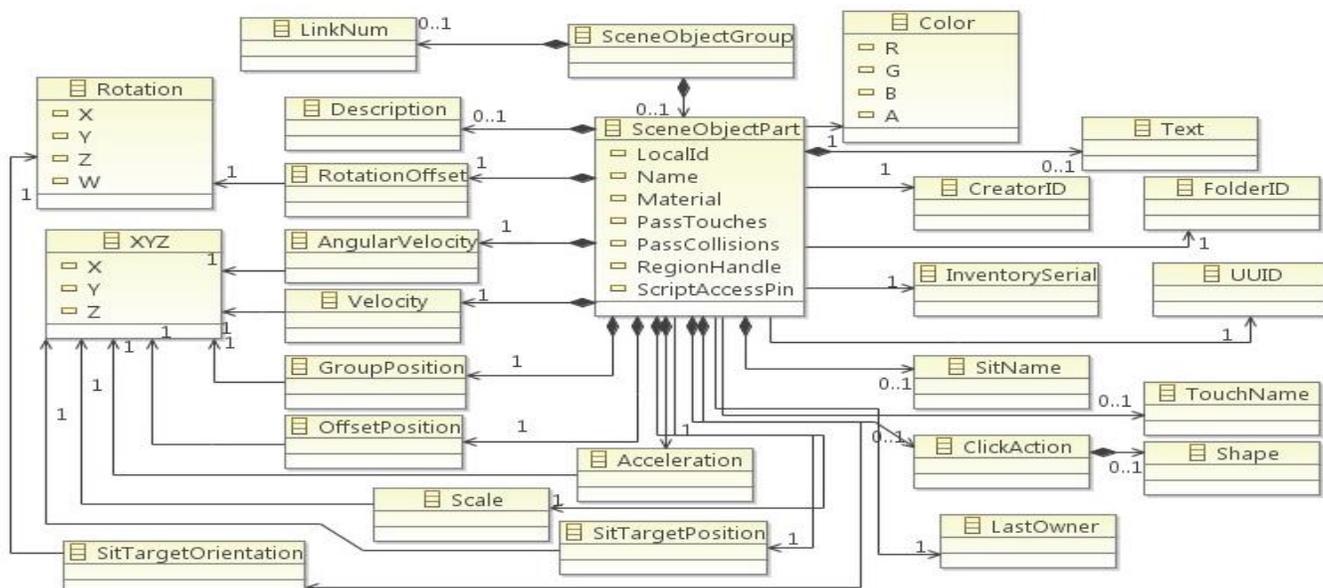


Fig. 9. Simplified Meta-Model of Object Presentaion in OpenSimulator Environment.

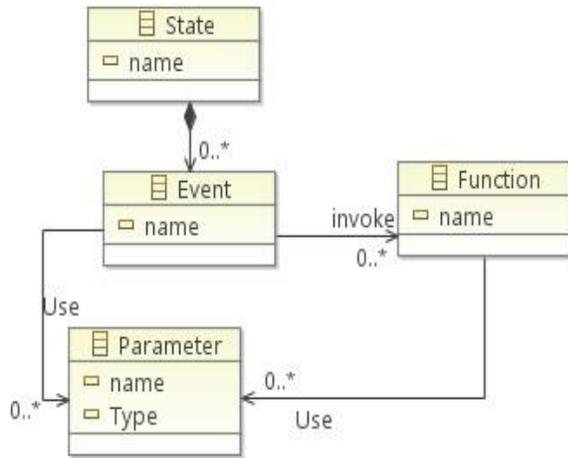


Fig. 10. Simplified Meta-Model of LSL Language Structure.

Linden Scripting Language is a state-event driven scripting language, a type of finite state machine. A script consists of variables, function definitions, and one or more named states. Each state contains a description of how to react to events that occur while the program is within that state. The system sends to the script, such as timers, movement, chat (from other agents), email, and collisions (with objects in the virtual world). Scripts can change most aspects of the state of the object and communicate with other objects and agents. As soon as a script is added to an object, and turned on, it begins to execute. The Fig. 10 shows a simplified LSL’s meta-model.

4) *Transformation rules*: The transition from XOS to OAR model requires firstly a correspondence between the XOS and OAR elements. This correspondence is divided into two parts: the first one is for the transformation of 3D objects and their positions on the scene (Table II); the second part is the transformation of different animation’s scripts (Table III).

TABLE II. XOS TO OAR OBJECT FORMAT

XOS	OAR objects format
Shell	SceneObjectGroup
Mesh	A file with .lmesh extension in the asset directory
Xyz	Xyz
Color	Color
Scale	Scale
Position	GroupPosition
Rotation	RotationOffset

TABLE III. PLUGIT TO LSL FUNCTION

Plugit	LSL
Action	Function
Param	Parameter
Event	Event

```

default
{
    touch(integer num_detected)
    {
        llSetPos(llGetPos() + <0,0,2>);
    }
}
    
```

Fig. 11. Example of Generated LSL Code.

Animations in XOS are described in XML format, and grouped into several plugins. Although the components of each plugit differ from one plugit to another (especially at the number of parameters) and depending on the desired animation, there is a standard part that describes the function of the animation and the event that triggers it. For example, the description of the plugin of moving an object by clicking on the left mouse button will be as follows:

```

<link action="Goto" instance="ball+goto+0%2e2"
group="Scene" event="LeftClick"/>
    
```

Each attribute has a meaning:

- Action: the name of the function to be called.
- Event: the name of the event that triggers the action.
- Instance: The name of the instance that contains the parameters needed to call the "GoTo" function, and which in turn, has a particular description containing the new position (x, y, z) of the object after executing the "Goto".

By respecting the transformation rules proposed in Table III, we obtain an LSL file containing the code of the LSL animation (Fig. 11):

The generated LSL code consists of several blocks. The first block is "default", which represents the default state of the object. An object can have several states depending on its nature, the person object for example can be on or sitting, the light object can be on or off. The second block "touch" represents the event that will change the position of the object (in our case it is the event click on the right button of the mouse). The "llsetPosition" function is the LSL function that changes the position of the object according to the cardinalities (x, y, z).

The generated script is placed in the "assets" folder in the OAR archive, and it is linked to the "ball" object via its unique name. Each object in "OAR" format has a unique identifier, and each script is attached to its object with the same name as the object in question.

IV. CASE STUDY

A. Presentation, Protocol, and Objective of the Workshop

Mechanics remain a special educational area, this comes from the fact that the majority of the notions, introduced in its courses, need visual experiments and practical activities so that a learner can assimilate these notions and understand their usefulness.



Fig. 12. Practical Activity with Real Equipment.

Mechanics remain a special educational area, this comes from the fact that the majority of the notions, introduced in its courses, need visual experiments and PAs so that a learner can assimilate these notions and understand their usefulness. The case presented in this section of the paper illustrates the study of a ball free fall where the ball is subject to a single force, its weight, and the other forces are neglected. The main objective is to determine the relationship between the ball's height and the free fall time. The learner takes a ball with fixed mass and tries to change the height in order to launch the ball. An electronic clock measures the travel time (t) of the ball from its origin. The learner repeats the experimentation many times by changing the height of the ball (h), notes the fall time on the clock, calculates and notes the quotient h/t^2 .

This experiment generally requires a hardware device called a free fall device (Fig. 12). The latter includes several parts, namely a digital counter for calculating the falling time, a micro-magnet which retains the ball in the starting position, a switch for starting the time measurement, and finally a plate on which the ball strikes in order to stop the time measurement.

B. Virtualization Steps

The virtualization of the workshop described above has undergone several steps. First, we have developed a list of minimal requirements from hardware devices. There are the ruler with which the learner can choose the height of the ball, a ball and a display showing the time of the free fall.

The second step consists in modeling the various devices in three dimensions. To do so, we used the "blender" software as shown in Fig. 13.

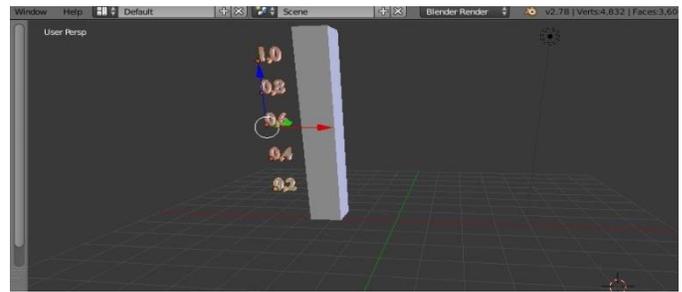


Fig. 13. Creation of Needed Objects in Blender Software.

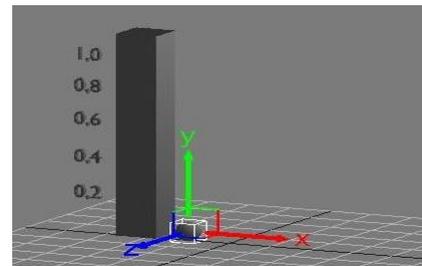


Fig. 14. Importation of All Objects in OpenSpace 3D Software.

Once the objects are created, Blender offers the possibility of exporting the scene to several formats. We chose "collada" format which is the standard format for exchanging 3D objects.

Step three consisted of programming the interaction of the virtual devices with the users in order to give them the opportunity to simulate the practical workshop and to collect information (the displayed time of the free fall). For this, we used the OpenSpace3D software. We have imported 3D objects and then created a scene. We placed the different objects in their initial positions and animate the objects so that if the learner clicks on a number on the ruler (a chosen height), the ball changes its position and moves to the opposite of the mentioned height, see (Fig. 14).

Once the user clicks on the ball, the ball falls freely to the ground and the time counter gives the duration of the fall. To ensure this interaction, we have called the various plugIts of the OpenSpace3D software (Fig. 15). The latter offers a variety of animation plugIts allowing the change of the position of an object, rotation, or even treating object collision without coding.

It remains to point out that the 3D engine used in OpenSpace3D offers the exploitation of physics laws by giving the possibility to parameterize gravity, mass and many other options by proposing a plugIt for that.

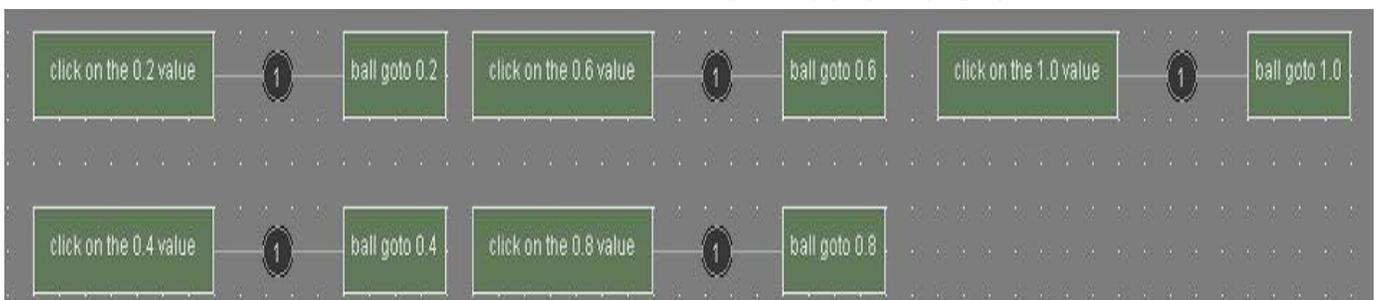


Fig. 15. Animating Objects in OpenSpce 3D Software.

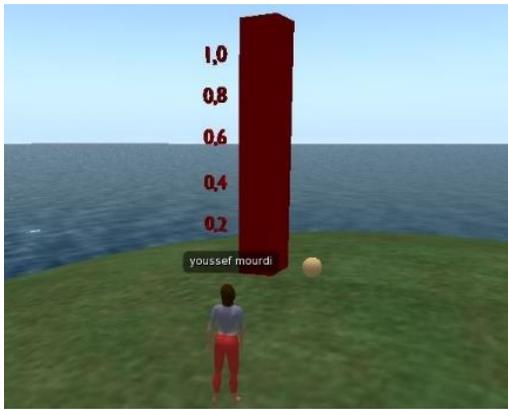


Fig. 16. Obtained View after OAR Importation.

The last step was to take the generated OpenSpace3D file with the XOS extension after setting up the interactions and to import it into our transformation module to get the final result, which is an OAR package containing the files, mesh, materials, and all the components of our PA in order to import it on the OpenSimulator platform. We notice that the objects are imported and placed correctly on the OpenSimulator scene.

C. Results

We have imported the resulting archive of the transformation module to OpenSimulator, using the commands of OpenSimulator console to load the OAR archive. The result is very satisfying in terms of the devices presentation (Fig. 16).

The first challenge was to import a scene of objects from an XOS description to an OAR structure. In accordance with the transformation rules described in Section 4, we obtained the following OAR structure (Fig. 17).

As regards to the layout of the objects on the scene, the transformation module was based on the coordinates (x, y, z) described in the XOS files. The transformation module imported the necessary meshes and generated an XML file for each object that described it in the "assets" folder of the OAR archive (Fig. 18).

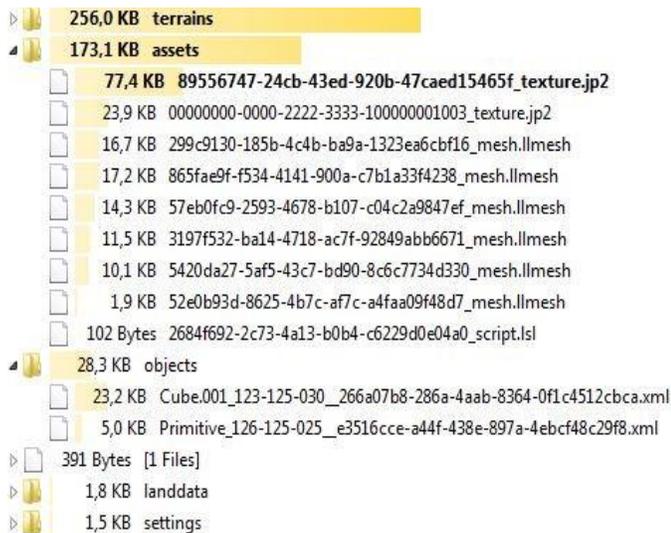


Fig. 17. The Generated OAR File.

```
<GroupPosition>
  <X>126.343</X>
  <Y>125</Y>
  <Z>29.08239</Z>
</GroupPosition>
<OffsetPosition>
  <X>0</X>
  <Y>0</Y>
  <Z>0</Z>
</OffsetPosition>
<RotationOffset>
  <X>0.2008292</X>
  <Y>0.677988</Y>
  <Z>-0.2008292</Z>
  <W>0.6779879</W>
</RotationOffset>
```

Fig. 18. Position of the Group Objects.

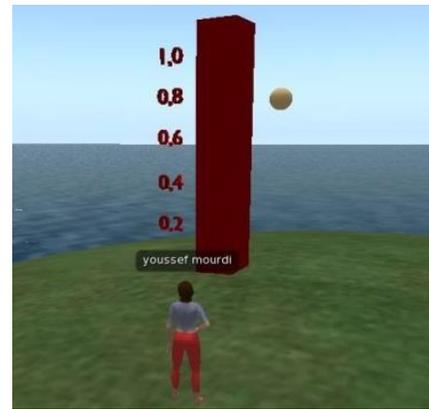


Fig. 19. Interaction between the Learner and the Object.

The second challenge we had was to keep the objects interacting with the users. To deal with this, the transformation module generated the necessary LSL files containing the calls to functions already programmed in LSL (Fig. 11).

When a user (learner) touches a number (selected height), the ball moves to the chosen height as seen in Fig. 19.

V. CONCLUSION

E-learning is a new field that has its roots in several sectors such as education and training, open distance learning, knowledge management, quality etc. National and international organizations, academic institutes and experts focus on analysing the success of e-learning and on how best to develop new approaches and structures to ensure ever better quality. The integration of practical activities is a very important task and is the cornerstone of a high quality education.

The practical activities of e-learning are progressing rapidly, taking advantage of the latest technologies and using various principles of pedagogy and organization. Today, their integration is characterized by a high level of innovation, experimentation and research of the most adapted solutions, such as remote control of real equipment, videos or virtual laboratories, that require more improvements because they do not really adapt to the expectations and needs of the learners or to the habits of teachers.

A common framework therefore needs to be put in place to both ensure a realistic open environment accessible to all and to provide learners with a high level of immersion while

performing their practical activities. This was the purpose of this paper where we give teachers, the opportunity to virtualize the practical activities, and to use the forces of virtual worlds in terms of communication, immersion, and the degree of realism for pedagogical purposes especially in the fields of distance learning and training.

The main objective of our research was to develop a transformation module allowing a user to import scenes of interactive objects from the OpenSpace3D software to a format recognized and importable in the virtual worlds of OpenSimulator. This transformation module helped in transforming a scene of animated objects using OpenSpace3D to an OAR archive importable on the virtual worlds of OpenSimulator. This proposal will allow teachers and especially non-computer scientists to create virtual workshops and make the various devices interactive without coding anything; this is thanks to OpenSpace3D software.

The results obtained in this paper are very satisfying. The limitations, which we must remedy later, are generally at the level of the transformation module development. The functions of OpenSimulator are not always the same adopted by OpenSpace3D in terms of signature and parameters. Consequently, future works will focus on improving the architecture proposed in this paper, as well as the search for integration of an intelligent interface to ensure the best transformations.

REFERENCES

- [1] K. Robins and F. Webster, "The Virtual University?," *Virtual Univ. Knowledge, Mark. Manag.*, pp. 3–19, 2002.
- [2] W. R. Watson and S. L. Watson, "What are learning management systems, what are they not, and what should they become?," *TechTrends*, vol. 51, no. 2, pp. 28–34, 2007.
- [3] L. D. Feisel and A. J. Rosa, "The role of the laboratory in staff development.," *J. Eng. Educ.*, vol. 94, no. 1, pp. 121–130, 2005.
- [4] C. Coti, J. V. Loddo, and E. Viennet, "Practical activities in network courses for MOOCs, SPOCs and eLearning with Marionnet," 2015 Int. Conf. Inf. Technol. Based High. Educ. Training, ITHET 2015, 2015.
- [5] M. Casini, D. Prattichizzo, and A. Vicino, "The automatic control telab: A remote control engineering laboratory," *Proc. 40th IEEE Conf. Decis. Control. Vols 1-5*, no. February, pp. 3242–3247, 2001.
- [6] C. Lazar and S. Carari, "A remote-control engineering laboratory," *IEEE Trans. Ind. Electron.*, vol. 55, no. 6, pp. 2368–2375, 2008.
- [7] D. Hercog, B. Gergic, S. Uran, and K. Jezernik, "A DSP-Based Remote Control Laboratory," *IEEE Trans. Ind. Electron.*, vol. 54, no. 6, pp. 3057–3068, 2007.
- [8] E. Guimarães et al., "REAL: A Virtual Laboratory for Mobile Robot Experiments.," *IEEE Trans. Educ.*, vol. 46, no. 1, p. 37, 2003.
- [9] L. Xu, D. Huang, and W. Tsai, "V-lab: a cloud-based virtual laboratory platform for hands-on networking courses," *Proc. 2012 ACM Conf. Innov. Technol. Comput. Sci. Educ.*, pp. 256–261, 2012.
- [10] J. Lima, L. Morgado, B. Fonseca, P. Martins, and H. Paredes, "Effectiveness of virtual world timers in educational physics simulations," *SLACTIONS 2011 - Res. Conf. Second Life world Life, imagination, Work using metaverse platforms*, November, 2011, 2011.
- [11] B. Dalgarno, "The potential of virtual laboratories for distance education science teaching: Reflections from the development and evaluation of a virtual chemistry laboratory," *Proc. Aust. Conf. Sci. Math. Educ. (formerly UniServe Sci. Conf.)*, vol. 9, pp. 90–115, 2012.
- [12] K. Andreas, T. Thrasyvoulos, D. Stavros, and P. Andreas, "Collaborative learning in OpenSim by utilizing Sloodle," in 6th Advanced International Conference on Telecommunications, AICT 2010, 2010, pp. 90–95.
- [13] T. Ritzema and B. Harris, "The Use of Second Life for Distance Education," *J. Comput. Sci. Coll.*, vol. 23, no. 6, pp. 110–116, 2008.
- [14] C. Allison et al., "Growing the Use of Virtual Worlds in Education: an OpenSim Perspective," *EiED 2012 Proc. 2nd Eur. Immersive Educ. Summit*, pp. 1–13, 2012.
- [15] M. Rico, J. Ramirez, D. Riofrio, M. Berrocal-Lobo, and A. De Antonio, "An architecture for virtual labs in engineering education," *IEEE Glob. Eng. Educ. Conf. EDUCON*, pp. 210–214, 2012.
- [16] M. J. Callaghan, K. McCusker, J. L. Losada, J. Harkin, and S. Wilson, "Using game-based learning in virtual worlds to teach electronic and electrical engineering," *IEEE Trans. Ind. Informatics*, vol. 9, no. 1, pp. 575–584, 2013.
- [17] "OpenSpace 3D." [Online]. Available: <http://www.openspace3d.com/>. [Accessed: 30-Jan-2017].
- [18] R. González Crespo, R. F. Escobar, L. Joyanes Aguilar, S. Velasco, and A. G. Castillo Sanz, "Use of ARIMA mathematical analysis to model the implementation of expert system courses by means of free software OpenSim and Sloodle platforms in virtual university campuses," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7381–7390, 2013.
- [19] M. Lahti, H. Hätönen, and M. Välimäki, "Impact of e-learning on nurses' and student nurses knowledge, skills, and satisfaction: A systematic review and meta-analysis," *Int. J. Nurs. Stud.*, vol. 51, no. 1, pp. 136–149, 2014.
- [20] R. Heradio, L. De La Torre, D. Galan, F. J. Cabrerizo, E. Herrera-Viedma, and S. Dormido, "Virtual and remote labs in education: A bibliometric analysis," *Comput. Educ.*, vol. 98, pp. 14–38, 2016.
- [21] C. A. Jara, F. A. Candelas, S. T. Puente, and F. Torres, "Hands-on experiences of undergraduate students in Automatics and Robotics using a virtual and remote laboratory," *Comput. Educ.*, vol. 57, no. 4, pp. 2451–2461, 2011.
- [22] E. Fabregas, G. Farias, S. Dormido-Canto, S. Dormido, and F. Esquembre, "Developing a remote laboratory for engineering education," *Comput. Educ.*, vol. 57, no. 2, pp. 1686–1697, 2011.
- [23] M. Stefanovic, V. Cvijetkovic, M. Matijevic, and V. Simic, "A LabVIEW-based remote laboratory experiments for control engineering education," *Comput. Appl. Eng. Educ.*, vol. 19, no. 3, pp. 539–549, 2011.
- [24] C. Terkowsky, C. Pleul, I. Jahnke, and A. E. Tekkaya, "Tele-operated laboratories for online production engineering education platform for e-learning and telemetric experimentation (PeTEX)," *Int. J. Online Eng.*, vol. 7, no. SUPPL., pp. 37–43, 2011.
- [25] V. Fernandez et al., "'Low-Cost Educational Videos' for Engineering Students: a new Concept based on Video Streaming and YouTube Channels," *Int. J. Eng. Educ.*, vol. 27, no. 3, pp. 518–527, 2011.
- [26] G. Quesnel, R. Duboz, and É. Ramat, "The Virtual Laboratory Environment - An operational framework for multi-modelling, simulation and analysis of complex dynamical systems," *Simul. Model. Pract. Theory*, vol. 17, no. 4, pp. 641–653, 2009.
- [27] Y. Daineko and V. Dmitriyev, "Software module 'Virtual Physics Laboratory' in Higher Education," in 8th IEEE International Conference on Application of Information and Communication Technologies, AICT 2014 - Conference Proceedings, 2014.
- [28] C. Tüysüz, "The effect of the virtual laboratory on students' achievement and attitude in chemistry," *Int. Online J. Educ. Sci.*, vol. 2, no. 1, pp. 37–53, 2010.
- [29] D. Raineri, "Virtual laboratories enhance traditional undergraduate biology laboratories," *Biochem. Mol. Biol. Educ.*, vol. 29, no. 4, pp. 160–162, 2001.
- [30] S. Diwakar, K. Achuthan, and P. Nedungadi, "Biotechnology virtual labs- integrating wet-lab techniques and theoretical learning for enhanced learning at universities," in DSDE 2010 - International Conference on Data Storage and Data Engineering, 2010, no. February, pp. 10–14.
- [31] S. Diwakar, K. Achuthan, P. Nedungadi, and B. Nair, "Biotechnology Virtual Labs : Facilitating." 2012.
- [32] T. Mohsen-Torabzadeh, Z. Muhammed Hossain, P. Fritzson, and T. Richter, "OMWeb - Virtual Web-based Remote Laboratory for

- Modelica in Engineering Courses,” in Proceedings of the 8th International Modelica Conference, 2011, pp. 153–159.
- [33] V. Potkonjak, M. Vukobratovi?, K. Jovanovi?, and M. Medenica, “Virtual Mechatronic/Robotic laboratory - A step further in distance learning,” *Comput. Educ.*, vol. 55, no. 2, pp. 465–475, 2010.
- [34] M. Stefanovic, “The objectives, architectures and effects of distance learning laboratories for industrial engineering education,” *Comput. Educ.*, vol. 69, no. September 2014, pp. 250–262, 2013.
- [35] J. Defazio, T. Faas, and R. Finch, “Building multi-user virtual worlds,” *Proc. CGAMES’2013 USA*, pp. 132–137, 2013.
- [36] D. J. Ketelhut, B. C. Nelson, J. Clarke, and C. Dede, “A multi-user virtual environment for building and assessing higher order inquiry skills in science,” *Br. J. Educ. Technol.*, vol. 41, no. 1, pp. 56–68, 2010.
- [37] A. Pinheiro et al., “Development of a mechanical maintenance training simulator in OpenSimulator for F-16 aircraft engines,” *Entertain. Comput.*, vol. 5, no. 4, pp. 347–355, 2014.
- [38] M. B. Ibáñez, J. J. García, S. Galán, D. Maroto, D. Morillo, and C. D. Kloos, “Multi-user 3D virtual environment for Spanish learning: A wonderland experience,” in Proceedings - 10th IEEE International Conference on Advanced Learning Technologies, ICALT 2010, 2010, pp. 455–457.
- [39] “Scol programming language.” [Online]. Available: http://redmine.scolring.org/projects/scol/wiki/Scol_Language_particular_syntax. [Accessed: 30-Oct-2017].
- [40] “OpenSimulator.” [Online]. Available: http://opensimulator.org/wiki/Main_Page. [Accessed: 30-Oct-2017].

Performance Evaluation of Cloud Computing Resources

Muhammad Sajjad*, Arshad Ali*, Ahmad Salman Khan†

*Department of Computer Science & Information Technology, The University of Lahore, Lahore, 55150, Pakistan

†Department of Software Engineering, The University of Lahore, Lahore, 55150, Pakistan

Abstract—Cloud computing is an emerging information technology which is rapidly growing. However, measuring the performance of cloud based applications in real environments is a challenging task for research as well as business community. In this work, we focused on Infrastructure as a Service (IaaS) facility of cloud computing. We made a performance evaluation of two renowned public and private cloud platforms. Several performance metrics such as integer, floating Point, GFLOPS, read, random Read, write, random write, bandwidth, jitter and throughput were used to analyze the performance of cloud resources. The motive of this analysis is to help cloud providers to adjust their data center parameters under different working conditions as well as cloud customers to monitor their hired resources. We analyzed and compared the performance of OpenStack and Windows Azure platforms by considering resources like CPU, memory, disk and network in a real cloud setup. In order to evaluate each feature, we used related benchmarks, for example, Geekbench & LINPACK for CPU performance, RAMspeed & STREAM for memory performance, IOzone for disk performance and Iperf for network performance. Our experimental results showed that the performance of both clouds is almost same; however, OpenStack seems to be better option as compared to Windows Azur keeping in view its cost as well as network performance.

Keywords—Cloud computing; OpenStack; windows azure

I. INTRODUCTION

Cloud Computing is an evolutionary technology which delivers computing software as well as hardware resources as a service over the internet. Cloud computers are interconnected and virtualized in a distributed and parallel system. The access to the infrastructure and computing resources such as compute, memory, storage, network, software, application and platform is permissible for any user for building applications.

The cloud services are mainly classified as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). IaaS provides infrastructural hardware, the most frequently used cloud, as a service to the users. The cloud infrastructure allows service providers to offer the infrastructural facility to the customers so that they can access the IaaS resources and enjoy the better service provision for smooth working of their applications without buying other resources. The clients just pay for the resources and services used on the basis of adapted Service Level Agreement (SLA) with a service provider [1]. The cloud clients pay for these resources just like as they pay for utility services such as water, gas, and electricity as per need and use [2].

Amzaon, Rackspace, Google, Eucalyptus, XEN, OpenNebula, Nimbus, Microsoft and OpenStack are major cloud computing resource/service providers. Monitoring of cloud resources is important both for cloud service providers and for cloud customers. Cloud providers monitor the efficiency and current status of assigned resources in order to handle future requests from customers. Monitoring helps cloud customers to investigate the resources assigned to them, and ensures that they get the demanded resources they are paying for. Further, it allows them to know when to demand for additional resources, when to surrender any underutilized resource, and what amount of numerous physical resources are suitable for different type of applications.

OpenStack is open source technology to provide elastic cloud operating systems. Windows Azure is public, private and hybrid cloud platform.

Moreover, diverse nature of different types of applications comes with different demands leading toward the need for different features in a platform naturally. Other aspects such as service models and prices are also taken into consideration. Performance evaluation of cloud services and resources is an important issue for cloud customers as well as for cloud providers. Cloud computing performance can be in terms of response time, throughput, reliability, security and availability. Monitoring of cloud services over internet based applications in a real cloud setup is much needed but difficult task. Research as well as business community is paying much attention to improve performance of IaaS resources.

This work mainly focused on performance analysis of two well-known IaaS clouds platforms, i.e., OpenStack [3] and Windows Azure [4] and compared their performance by considering various cloud resources such as CPU, memory, disk and network. OpenStack is a popular and fast growing open source cloud computing for private, public and hybrid clouds while Windows Azure is one of the mostly used private clouds. Our results for comparison analysis are based on configuration made in a real time environment.

A. Objectives/Contribution

The major objective of this work is to evaluate and compare performance of various IaaS resources of two famous cloud platforms i.e., OpenStack & Windows Azure platform in a real cloud environment. The considered IaaS resources include CPU, memory, disk and network.

B. Research Approach

The research process for this work comprised of the following steps:

- Survey and literature review about performance evaluation of IaaS resources of various cloud platforms.
- Cloud platform selection included two clouds OpenStack and Windows azure
- Setup an infrastructure
 - We used two servers HP ProLiant DL380 G7 with same specification of resources like CPU, RAM, Hard disk and Network.
 - For OpenStack Cloud, we first installed Ubuntu 14.04 operating system, and then Juno release of OpenStack was installed and configured. After completing the configuration, different tasks were performed for testing purpose.
 - For Windows Azure Cloud, windows Server 2012 R2 operating system was installed, then Hyper-V Role added for virtualization. Then, Domain controller, Microsoft SQL server 2012 R2, Microsoft Virtual Machine Manager 2012 R2, Service Provider Foundation, System Center 2012 R2 Orchestrator, Windows Assessment and Deployment Kit 8.1, Windows Azure pack were installed and configured. Once the configuration was successful, then some service accounts were created to run Azure cloud services.
- Benchmarking the cloud based on benchmarking of CPU performance, Memory performance, Disk performance and network performance for evaluation for determining the performance in real machines.
- Different benchmarking tools are available to test performance of cloud resources. Among the available benchmarking tools, we used Geekbench, LINPACK, RAMspeed, STREAM, IOzone and Iperf benchmark tools depending upon the nature of cloud resource being evaluated.
- Next step was to select parameters for performance evaluation of cloud resources like CPU, Memory, Disk and Network in real time environment. The selected parameters for each resource were as follows:
 - CPU Performance: Minimum (GFLOPS), Maximum (GFLOPS), Average (GFLOPS), Integer (Single core/Multicore) and Floating point (Single core/ Multicore)
 - Memory Performance: Integer (Average), Floating Point (Average)
 - Disk Performance: Read, Random Read, Write, Random write
 - Network Performance: Bandwidth, Jitter and Throughput

- Results were obtained for parameters selected for each resource by using related benchmarking tools. Average results were calculated by repeating the process a number of times and repeated a number of times to drive rational conclusion.
- The comparative results were presented in the form of graphs by using MATLAB.

C. Organization of the Paper

The rest of the paper is organized as follows: Section II provides detailed background of the problem under study and related works in this regard. In Section III, we provide the experimental set up. Section IV presents benchmarking tools used for performance evaluation and comparison. In Section V, results obtained through experiments and related observations are provided. Finally, Section VI concludes the work and hints towards further research issues and challenges.

II. BACKGROUND AND RELATED WORK ABOUT PERFORMNACE OF CLOUDS

Performance of cloud resources is very crucial for cclients as well as for service providers. Consequently, such performance evaluation is essential from the point of view of cloud service providers and clients.

A. Background

Cloud service providers are interested in evaluating the performance of different infrastructure based cloud resources such as compute, storage, network and virtual machines. A single individual component is unable to provide complete performance report of a cloud. Infrastructure Response Time (IRT) is a new approach to get more accurate performance of virtualized cloud environment. IRT is explained as the time it takes for an application to put a request (I/O) over virtual environment and get back its response. This request can be a normal data transfer between two Virtual Machines or complex one like transaction of database and storage into a storage array.

The most common idea for achieving best performance is by increasing the resources. However, customers are forced to bear higher costs when they opt for purchasing their own resources which is not a better solution. Therefore, cloud customers use needed resources from cloud service providers and pay as per use without investing heavy amount in infrastructure setup. Therefore, cloud consumer is interested in performance of applications hosted on the cloud platform which greatly depends upon performance of IaaS resources. Application Response Time (ART) is an important metric in application performance management which is calculated as time taken by the application to respond to other users' requests. Thus, cloud provider is more interested to have a complete view of health of whole cloud for better service provisioning [5].

In a cloud environment, the tasks can be classified as computation intensive and communication intensive tasks. RAM and CPU cores are important cloud resources consumed by computation-intensive tasks. In a cloud application, a communication-intensive task normally produces large number of network transactions between cloud user devices and cloud

systems. Therefore, network monitoring is critical to analyze the estimation of network performance monitoring (NPM) which is analyzed through different network monitoring techniques. Active monitoring [6,7], method of using SNMP agent, and passive monitoring [8] are normally three strategies for network monitoring.

Computation-intensive tasks, such tasks can be divided into CPU-intensive tasks and memory-intensive tasks [9]. For energy efficiency of cloud systems, Lefebvre and Orgerie [10] analyzed multicore platform with focus on CPU cores only. This work provided an assessment of the energy consumption during relocation of VMs only with computation-intensive cloud applications. Liu et al. [11] provided a new cloud infrastructure which can dynamically associate Virtual Machines (VMs) based on CPU utilization of servers to detect idle physical servers. Idle physical servers can be turned off to protect energy. However, these energy protecting strategies do not take into account the work load in cloud systems and hence are very coarse-grained.

In communication intensive tasks, quantity of user's requests and related data size can have huge impact on the system performance [12]. Many web services are installed inside servers (weather physical or virtual). If webhosting server serves the peaks, then it is most likely over-provisioning when the demand is high. In cloud computing, the performance analysis of network resource is necessary for best use of resources as in traditional computing. Cloud infrastructure consists of different component like server virtual machine, network interface and users are required to select component according to need. A well-known communication intensive application for cloud computing is Community Atmospheric Model (CAM) [13], a massive parallel application used in worldwide weather prediction.

B. Related Work about Performance

The objective of performance evaluation of cloud computing is to investigate and compare the performance attributes of the system [14]. Amazon Elastic Cloud (EC2), Microsoft Windows Azure, GoogleApps Engine, Sales-force and Drop-box are well known commercial cloud service providers. Commercial success of any cloud computing platform depends upon its ability to deliver guaranteed Quality of Services (QoS) [15]. Performance can be evaluated through measurements, simulation and/or modeling [16].

Zhang et al. [17] performed an evaluation of four commercial cloud platforms and drew a conclusion that prevailing platforms provide different types of services which are offered at various levels of abstraction. Therefore, end-users should select more than one cloud platform keeping in view their requirements to satisfy business needs.

Buyya et al. [18] provided an evaluation of some cloud platforms considering market oriented cloud computing. The study focused on the need for advent of techniques for business cloud management on the basis of risk management and customer's requirements.

Hofer et al. [19] provided classification by considering features of various cloud systems. This work considered various characteristics which include service model, license type, cost model, supported languages and operating systems, virtualization mechanism and development tools.

Rimal et al. [20] made a classification on the basis of features and used it for the purpose of comparison of system offered by providers. The attributes of providers considered for comparison were architecture, interoperability and security, virtualization technique; services provided mechanisms for load balancing, and support for software and programming languages.

Ostermann et al. [21] analyzed the performance of EC2 utilizing micro-benchmarks like LMbench, Bonnie, CacheBench, HPC Challenge (HPCC) and kernels. They concluded that achieved performance of virtualized resources from public clouds is lower when compared with the theoretical performance limits, particularly for compute and network intensive applications. They compared the observed virtualized performance of a private cloud with the non-virtualized performance. They used metrics such as CPU, I/O, and memory hierarchy on the Single Instance benchmarks to evaluate performance.

Some studies performed quantitative comparisons among different providers. For example, authors of [22] proposed a framework namely Cloud Comp which provides performance comparison of various providers [22].

Zheng et al. [23] carried out comparison of four commercial cloud providers, namely Amazon EC2, Google AppEngine, Windows Azure and Rackspace cloud servers on the basis of a few components of computing, network, database and storage. Different issues affecting startup time of cloud VMs across Amazon EC2, Google AppsEngine, Windows Azure and Rackspace are studied in [24].

Li et al. [25] carried out a performance and cost comparison between four major public clouds. The clouds are compared on the common functionality set, which incorporates elastic computing, intra-cloud network, persistent storage and Wide-range network.

Table I shows different studies reported on the topic. It classifies these studies as qualitative, taxonomy based and qualitative comparisons.

The latest research of cost effective cloud computing mostly analyzes the cloud service provider cost for offering cloud solution, for example, the authors in [26,27] explored cloud energy consumption and cost efficiency and discussed about different issues and challenges. According to [3], with on-demand resource provisioning and utility based costing; cloud service provider can really expand resource utilization and reduce their operational cost.

TABLE I. RELATED WORKS

Reference	Objective	Providers compared	Features compared	Comparison Type
Radu et al. [28]	Performance Evaluation of Azure and Nimbus Clouds for Scientific Applications	Microsoft Windows Azure Nimbus Cloud	Performance, computation speed, variability and cost models	Quantitative
Ang Li et al. [29]	CloudCmp: Comparing Public Cloud Providers	Amazon AWS, Windows Azure, Rackspace and Google AppEngine	Scaling latency, Operation response time, Time to consistency, Cost per operation, Response time, Throughput,	Simulation
Zhang et al. [17]	Cloud computing: state-of-the-art and research challenges	Amazon EC2 MWA Google AppEngine (GA)	Services, applications, virtualization mechanism and scalability	No Implementation. Theoretical work
Konstantinos et al. [30]	Comparison Between OpenNebula and OpenStack	OpenNebula and OpenStack	deployment time, migration time	Simulation
S. Itnal et Al. [31]	Network performance analysis and optimization on cloud	OpenStack	Network	No results Provided
Rimal et al. [20]	A Taxonomy and survey of cloud computing systems	AWS, WinAzure, GoGrid , SunCloud, Salesforce.com	Virtualization mechanism, services, reliability, interoperability	Taxonomy & Survey. No Implementation
G. V. Laszewski et al. [32]	Comparison of Multiple Cloud Frameworks	Nimbus, Eucalyptus, OpenStack, and OpenNebula	Software deployment, Interfaces, Storage, Networking, Hypervisors	Real Time
Hofer et al. [19]	Taxonomy of cloud computing services	Windows Azure Google ApEngine	Services, license model, QoS, payment model, security, standard etc.	Taxonomy-based No Implementation
R Ledyayev et al. [33]	High Performance Computing in a Cloud Using OpenStack	OpenStack	CPU, memory, network performance	Simulator
Li et al. [23]	A Factor Framework for Experimental Design for Performance Evaluation of Commercial Cloud Services	Amazon, GoGrid, Google, IBM, Microsoft, and Rackspace	Computation, Memory, storage , network, VM Instance	Working on factor framework No Implementation
Zach Hill et al. [4]	Early Observations on the Performance of Windows Azure	Windows Azure	virtual machines, storage services	Real Time Implementation
This work	Performance Analysis and Evaluation of Cloud Computing Resources	OpenStack Windows Azure	CPU, memory, disk and network	Real Time Implementation

C. Issues with Existing Approaches

Despite the fact that research community is focusing a lot on performance of cloud resources, however, in-depth knowledge about performance in a real set up is still lacking in single document. Most of the works do not provide true picture for performance analysis as those works are carried out using different types of simulators. Therefore, there is a need of having results and findings in a real set up instead of the same in a simulation environment.

D. Our Approach

While taking into consideration the related work about cloud performance analysis, we observed that (i) there is no

comparative study about performance evaluation of OpenStack and Windows Azure, and (ii) no in depth performance analysis related to selected cloud resources is found for analysis purpose.

To the best of our knowledge, this work is first of its in kind which provides detailed performance analysis of various critical issues including the price and benefits with applications in a cloud. We developed a real time environment to analyze the cloud performance like compute, network, and storage and disk workload. Here is some description which shows our work is unique.

First, we explore the performance of cloud computing and setup experiment by using OpenStack and Windows Azure.

Then, we verify our results by repeating experiments multiple times. The results of experiments can be helpful for cloud service providers in managing their services effectively in order to meet consumer's requirement. Moreover, our findings can provide useful insights to cloud consumers to manage idle resources more efficiently for smooth running of applications. The next section describes the experimental set up in detail.

III. EXPERIMENTAL SETUP

The experimental settings consist of following specification, we use two servers ProLiant DL380 G7 for cloud set up separately (refer to Table II for specifications of servers) and minimum requirement for client machine is Pentium 4 Machine with 512 MB RAM, network card,80GB Hard drive.

We installed OpenStack under Ubuntu on standalone machine and Windows Azure on a separate machine. Subsequent sub-sections describe the considered cloud platforms i.e. OpenStack and WinAzur.

A. OpenStack

In our experiments, we used the Juno version of OpenStack which is an open source software. Juno is the tenth release for structure of public, private, and hybrid clouds. The OpenStack contains lot of features to support application development. Many organizations like Rackspace, NASA, Citrix, Dell, Cisco, Canonical and many more participants of worldwide software community support OpenStack.

The OpenStack cloud allows service providers to propose computing resources by catering huge networks of VMs. To make an effective image provisioning, OpenStack stores images on the computes nodes, eliminating the needs of shifting the VM image on the network every time it is requested.

The hardware used for OpenStack implementation scenario is a HP ProLiant DL380 G7 series. This server has intel Xeon CPU E5620 dual processors and had a specific role that require huge processing capability. Server has the following specifications:

- 2 x Intel(R) Xeon (R) CPU E5620 @ 2.4GHz
- Intel chipset 5520
- 8 Cores
- 16 GB Memory
- 4 x 1Gb NIC's
- 3 x 300Gb SAS HDD (RAID-5)

Latest version of stable release of OpenStack from OpenStack repositories was downloaded and only one machine with above specification was used for our experimental. To assist flexible services, we used different OpenStack components for installation of cloud. The cloud controller and nova compute are installed on same server. Some additional software like MySQL data base server, RabbitMQ messaging queuing, Apache webserver and KVM are also installed on this machine.

TABLE II. SPECIFICATIONS OF SERVERS

Characteristic	Specification
Server Model	ProLiant DL380 G7
Chipset	Intel 5520
Processor type	Intel Xeon CPU E5620
System Architecture	64- bit
Processor speed	2.4GHz
Cores	8
No. of Processor	2
Main memory	16 GB
Network interconnect	Cisco 3560
Ethernet	1x4 GB
Network topology	Cisco proprietary
Virtualization	Yes
Hypervisor	Yes

Operating system
OpenStack Juno, Ubuntu 14.04, Microsoft Virtual Machine Manager 2012 R2, Service Provider Foundation, System Center 2012 R2 Orchestrator, Windows Assessment and Deployment Kit 8.1, Microsoft SQL server 2012 R2, Hyper-V, Windows Azure pack, and necessary tools for monitoring, Windows server 2012 Standard/Data Center Edition

B. WinAzur

WinAzur is a collection of integrated services like computing, storage, database, mobile and networking. Microsoft virtualization platform Hyper-V helps optimize hardware resources by combining multiple client operation systems on a single server. This describes the method of setting up a Private Cloud using Microsoft technologies such as Hyper-V & System center. The System Center delivers the fabric management and monitoring that is required for the services. Once the installation and configuration are complete, it is possible to use Microsoft System Center Virtual Machine Manager 2012R2 for a private cloud to be built and managed.

The Microsoft System Center allows cloud administration and management to deploy, monitor and report about it. The basic understanding about the roles and services in Windows 2012 R2, information of how to install SQL Server 2012 R2, The System Center Orchestrator 2012 R2 allows you to install the Service Provider Foundation and a practical knowledge System Center Virtual Machine Manager working. It is more necessary for a production environment, that careful installation of SQL server 2012R2 is required to ensure the proper working of System Center.

The service and workload layer discloses the knowledge of the Windows Azure Pack to Windows Azure. The WAP is built in a way that allows you to offer more services into the Windows Azure Pack.

The Service Management API is an application of the Windows Azure Service Management that provides a reliable customer API that talk to the WAP fabric underneath. If the services in the WAP are not deployed similar as in Windows Azure, the access to the Service Management API is consistent. The portal in the WAP looks same as the portal in Windows Azure. However, you are able to adjust the portal as you like as a service provider or if you don't like the portal at all, you can

build customized portal and use the required Management Service APIs. The provider portal allows the WAP administrator to design the WAP infrastructure to define plans that is assigned to an end user [34].

Windows instances are accessed through remote desktop to clients via public or private IP address, the password is encrypted and portal can be accessed through secure browser by using https. The server with same specifications as for OpenStack was used.

In windows azure virtual machines can be create, delete or re-create on demand and these virtual machines can be access just like physical server. To create virtual machine virtual hard disks (.vhdx files) are used. This supports image and disk categories of virtual hard disk. To create a virtual machine from image the following procedure can be used:

- Use Azure portal to make a virtual machine from image.
- Build and store a .vhdx file that comprises an image to WinAzure portal which is used to make a virtual machine.

Windows Azure gives particular arrangements of central processing unit (CPU) cores and memory for IaaS virtual machines. To create a virtual machine, select a particular size available from list however the size can be altered after deployment. The extreme size of a working disk can be 127 GB. When an OS disk is generated in Windows Azure three replicas of the disk are made for high availability.

Virtual machine storage: In Windows Azure a virtual machine is generated from an image or a disk. These virtual machines run one or various data disks installed operating system. All images are generated from virtual hard disk (.vhdx) which are stored in the form of blobs in in Windows Azure storage account.

Virtual machine network: The virtual machine systems are committed to virtual machine LAN activity. A virtual machine network can be two or more 1 GbE networks have been made through NIC Teaming.

IV. BENCHMARKS AND IMPLEMENTATION

The best approach to study the performance of specific system is to run the actual workload on the hardware platform and analyze its results. However, in certain scenarios, this method is not feasible. In such situations, analysts prefer to use typical benchmark results. Benchmarking is the basic technique for determining the performance of a real machine. Benchmarking mentions to running a set of typical programs on different workstations, networks and evaluating the results. Benchmark outcomes are used to analyze the performance of a particular system on agreed workload. Normally, comparative study of products depends upon benchmarks. They are used similar to monitoring and analysis software. System vendors, developers, and customers run benchmarks to identify performance problems of new systems.

In this research work, our concern was time and rate efficient evaluation of a particular cloud platform, as organizations and individuals usually do not want to spend too

much time and money on such concerns. We have chosen those benchmarks and platforms that strengthen this viewpoint.

Our focus was on actual hardware components for such as CPU, memory, disk, and network etc. Cloud providers are required to provide clear information such metrics.

A. Cloud Platform Selection

OpenStack cloud was our first choice due to the fact that it is mostly used in both industry and academia and it offers wide variety of service. The second choice was Microsoft windows Azure cloud, a new competitor in the IaaS platform, Microsoft traditional grip of the enterprise marketplace makes them a rational selection for various industries. In this research all VMs use the same version of Windows Server 2012 R2 64-bit operating system to reduce uncertainty and all instances on the particular platforms for the Windows instances. Furthermore, the chosen instance sizes and types are provided in Table III.

TABLE III. OPENSTACK AND WINAZUR MACHINE INSTANCES AND RESOURCES

VM Type		Resources (Same for both clouds)		
OpenSatck	WinAzur	VCPUS	RAM (MB)	Disk (GB)
m1.tiny	Micro	1	1024	20
m1.small	Mini	1	2048	20
m1.medium	Medium	2	4096	40
m1.large	High	4	8192	80

B. Benchmarks Selection for Performance Evaluation

In Cloud computing different applications have different hardware requirements. Our focus was to select those benchmarks that can evaluate cloud system resources like CPU, memory performance, storage and network. These parameters are appropriate to measure system's performance and the set of chosen benchmarks are applicable for an extensive mainstream of applications. We selected freely available and commonly used benchmarks which offer transparency, availability, and efficiency. Authors of [35] discussed a distributed testing model.

CPU performance: CPU performance is determined by two renowned parameters i.e. MIPS (Million Instructions Per Second) and FLOPS (Floating Point Operations Per Second). MIPS unit is tough to compare between CPU architectures, and workloads. CPU performance is discussed in [36].

Integer and Floating Point are computing intensive calculations. Integer data contains complete numbers, text and other similar items. But Floating Point Unit (FPU) procedures are further complex than integer. Examples of applications building a full usage of FPU are worksheets, graphical theory applications, games, and subsequently. Therefore, to achieve best performance processors required to perform the process of integer and FPU as fast as likely.

FLOPS are normally used to analyze the performance of a processor. A FLOPS simply determines floating point calculations and not integer operations. Therefore, FLOPS can

exactly calculate a processor floating point unit (FPU). In order to exactly measure the processing capabilities of a CPU, different types of tests needs to be run. For our analysis w.r.t. CPU performance, we used FLOPS, Integer and floating point.

For FLOPS, we selected **LINPACK** benchmark, a tool for performing numerical linear algebra. In this specific test we used LINPACK version developed by the Intel Corporation, which is open-source but with a closed source front-end called IntelBurnTest which is commonly used in overclocking circles.

To determine the performance of integer and floating point unit, **Geekbench** benchmark tool was chosen which is designed to work on different platform and it is used on Linux, Windows and Solaris.

Memory Performance: The metrics used for determining system performance in terms of memory are pretty simpler than CPU metrics as the memory system is debatably a simpler component than the CPU. RAMspeed and STREAM are renowned and most commonly used tools for benchmarking cloud performance w.r.t. memory.

Disk performance: The metrics used for calculating disk bandwidth are less complex and we are more anxious how rapidly the disk can read and write blocks of data to and from the hard drive. IOzone is a filesystem benchmark tool that is useful for performing an extensive file system analysis of a computer system. It allows number of options to be set as well as it contains a throughput choice in which you are required to mention limited parameters. We used IOzone method for quick evaluation of system's performance.

Network performance: The performance of network is decreased as a result of large VM working on the same machines. A large amount of data is communicated when a large number of machines use the network at the same time. A cloud platform needs fast network equipment and links in order to provide best performance. Bandwidth, latency, packet loss and jitter are interesting parameters concerning the network of a cloud provider. In this research we tried to evaluate the internal bandwidth of cloud providers by setting up instances within the same area and running network experiments between them. In order to determine the performance, we used Iperf tool which works as clinet-server model. Iperf, software for network analysis, can generate TCP and UDP data streams and calculate throughput of the network. It permits the user to set different parameters for analysis the network but more significantly one can assess bandwidth, jitter and throughput easily by modifying just a few of the typical configuration useful for quick evaluation.

V. RESULTS AND DISCUSSION

This section provides performance results of two clouds by using benchmarks described earlier. During experiments, it was ensured that no customer applications were in operation concurrently with the benchmarking applications on the experimental VMs. The only running processes were those which were essential for operating system to work itself at boot time. The subsequent sub-sections provide results with respect to performance of CPU, memory, disk and network.

A. CPU Performance

To measure the performances of the CPU, we used Geekbench and LINPACK tools. We used one Windows instance initially at a time to perform the performance tests on the cloud. The launched instances were running at 100% of CPU usage. The workload of Geekbench is divided into Integer Performance and Floating point performance. These workload measure the performance by performing intensive tasks related to processor that use heavily integer and floating-point instructions. High score lead to better overall performance. Floating point measures are important, for example in video games.

The benchmark conducted using the program IntelBurnTest with a matrix size of the Standard 512MB on m1. tiny and MicroVM instances while rest of the instances with standard 1024 MB, this test run 10 times on every machine being the default configuration. The complete outcomes of the test are shown in Tables IV, V and Fig. 1 to 3. The data using LINPACK benchmark shows Windows Azure gradually outperforms OpenStack (Fig. 1 to 3). OpenStack tiny & small instance outperforms Windows azure but Windows Azure medium & high instance outperforms OpenStack. The performance of windows Azure is a bit better than OpenStack in large instance.

The results shown in Fig. 1 are taken by using LINPACK tool while that shown in Fig. 2 and 3 are obtained by using Geekbench tool. Windows Azure performance in integer is some better than OpenStack. In integer all instances of Windows Azure outperform to all instance of OpenStack in single core as well as multicore integer.

TABLE IV. OPENSTACK CPU PERFORMANCE BY INSTANCE TYPE

LINPACK (in GLOPS)	LINPACK			Geekbench (Single Core/Multicore)	
	Min	Max	Average	Integer	Floating
m1.tiny	8.99	9.45	9.15	2068/2065	1953/1939
m1.small	9.44	9.67	9.61	2079/2052	1941/1955
m1.medium	16.02	18.00	17.50	1969/3824	1786/3529
m1.large	35.17	35.82	35.46	1996/7015	1842/7027

TABLE V. WINAZURE MEMORY PERFORMANCE BY INSTANCE TYPE

LINPACK (in GLOPS)	LINPACK			Geekbench (Single Core/Multicore)	
	Min	Max	Average	Integer	Floating
m1.tiny	8.79	9.24	9.04	2076 /2074	1971/ 1935
m1.small	9.21	9.64	9.48	2081/2091	1925/1969
m1.medium	18.75	18.97	18.88	2028/3827	1931/3861
m1.large	33.25	36.57	35.47	2000/7058	1857/6603

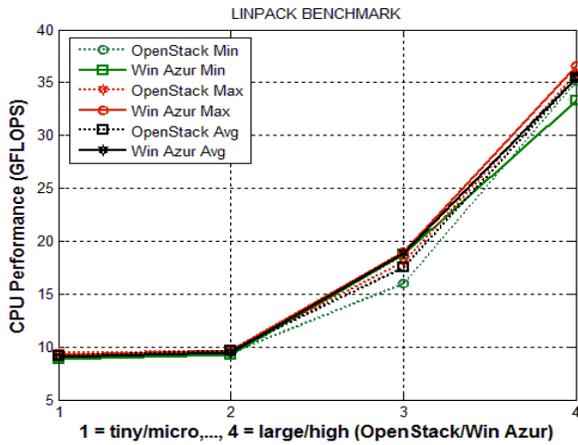


Fig. 1. CPU Performance Comparison (GFLOPS).

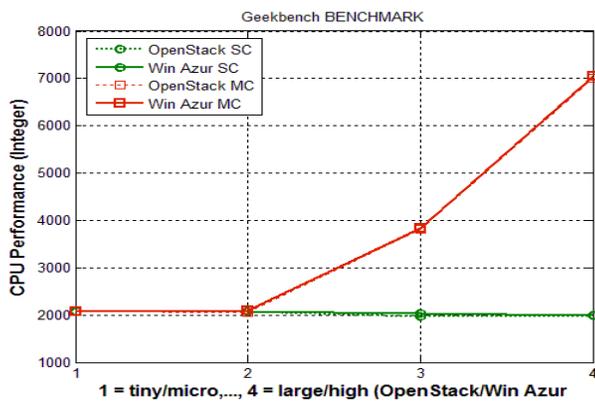


Fig. 2. CPU Performance Comparison (Integer).

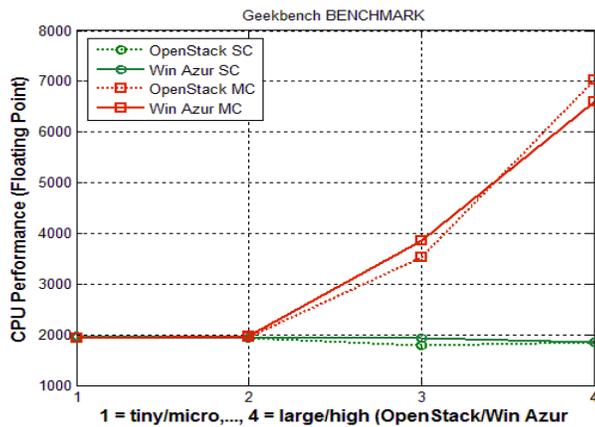


Fig. 3. CPU Performance Comparison (Floating Point).

This data shows two opinions of interest. The primary is that OpenStack gradually outperforms Windows Azure while Azure performance in floating Point is a bit better than OpenStack. In floating point, CPU performance for single core instance of Windows Azure is better, while in multicore OpenStack tiny and large instances perform better than that of windows Azure micro and high instances. Moreover, windows Azure in mini and medium instances outperforms in small and medium instances of OpenStack. Why this is so, is difficult to

say without more information about the particular datacenters. Moreover, both platforms scale in a way constant with the Cloud provider’s assurance in terms of CPU capacity. OpenStack for instance state that their tiny, small, medium and large instances each provide 1, 2 and 4 cores. In other way the CPU performance is almost twice every time the instance type is increased by a step. Microsoft states similar kind of facts about their platform. In summary, though both perform almost at the same level, but OpenStack, being feely available, seems to be a better option.

B. Memory Performance

We used RAMspeed and STREAM benchmark for memory performance. The performance of memory measured RAMspeed by using integer and floating point numbers. The purpose for using Integer and floating point is that it has to provide a lower limit on the memory performance as integer procedures are usually faster. Moreover, most of the mainstream applications use both integer and floating point numbers in which the floating point memory procedures be the bottleneck as far as memory goes.

The memory performance benchmark based on a bandwidth assessment as this is differentiates among System memories categories. Stream is memory benchmark that efforts to make best use of memory bandwidth, it is severely load the memory without using appropriate pressure on the CPU.

The command to measure memory performance using Integer with RAMspeed is `>ramspeed-win32 -b 3 -l 50` while the command to measure memory performance using Floating Point RAMspeed is `>ramspeed-win32 -b 6 -l 50`.

TABLE VI. OPENSTACK MEMORY PERFORMANCE BY INSTANCE TYPE

	RAMspeed (MB/s)		STREAM (MB/s)
	Integer (Average)	Floating Point (Average)	Average
m1. Tiny	7800.54	8417.12	8871.57
m1. Small	7828.22	8534.52	8968.45
m1. Medium	7693.76	8381.93	9282.34
m1. Large	7850.23	8437.52	9474.77

TABLE VII. WINAZURE MEMORY PERFORMANCE BY INSTANCE TYPE

	RAMspeed (MB/s)		STREAM (MB/s)
	Integer (Average)	Floating point (Average)	Average
Micro VM	7953.15	8576.95	9084.63
Mini VM	7691.57	8463.91	9274.47
Medium VM	7806.88	8502.19	9303.10
High VM	7654.17	8337.48	9498.42

The parameter -b is for bandwidth, 3 for Integer memory and 6 for Float memory, -l for length of benchmark we use 50 test and take average of them. The outcomes obtained by running RAMspeed and STREAM for both clouds are provided in Tables VI, VII and Fig. 4 and 5.

In RAMspeed integer performance in windows Azure’s micro and medium instances performance is better than OpenStack’s tiny and medium instances but OpenStack small and high instance performance is better than windows Azure mini and high instance (Fig. 4).

In STREAM benchmark windows Azure’s all instance performance is better than OpenStack all instances (Fig. 5). Windows Azure memory performance is much better than in OpenStack in terms of raw performance. Moreover, OpenStack shows a much less diverse performance than Windows Azure. In any case which metric worth most, consistent, or high performance based on application and condition. If experiments are run for a certain application and the outcomes show high difference but also that the real performance did not below the minimum definite point for the application to function correctly, then difference in the situation is not a problem. Difference only develops a problem if it differs with goes below that definite point. You need to choose in the event that lowest value that has most prominent impact on your application and select from subsequently. Finally, these values confirm that there is a lack of performance segregation in the cloud today.

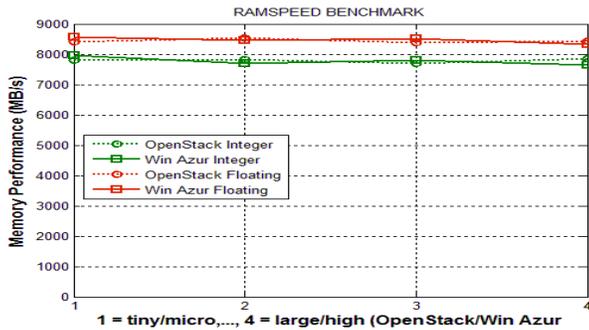


Fig. 4. Memory Performance Comparison using RAM Speed.

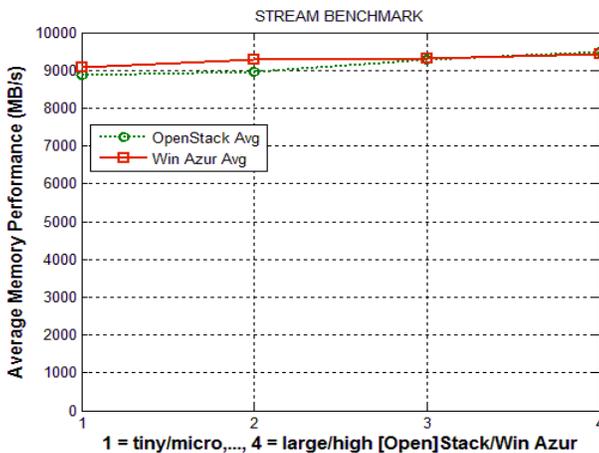


Fig. 5. Memory Performance Comparison using STREAM.

C. Disk Performance

Disk performance test is conducted on machines installed Windows operating system, the NTFS files system is used for benchmarking on the local drives. The theme of these tests is to calculate the time it proceeds and access data on the disk. We are interested in the entire sequence from main memory to disk and the other way from disk to main memory. In simple words we required the throughput speed. The software needs to know how many threads to use for the throughput test which was provided to it by initial testing the number of threads the CPU supports. The outcome obtained by using IOzone in throughput approach with some custom configuration indicating that we are concerned in sequential and random read and write speeds.

```
>iozone -i 0 -i 1 -i 2 -t 2
```

The 0=write/rewrite, 1= read/re-read, 2= random-read/write where -t flag indicate how many threads to use for the throughput test. We use 2 processes for throughput test conducted with each process writes a 512 Kbyte file in 4 Kbyte records and we take maximum throughput of per process in this benchmark.

The both provider’s information similar variance and the typical performance on both platforms are very similar as given in Tables VIII, IX and Fig. 6 and 7.

Read, Random read, Write and Random write of OpenStack Tiny, medium and large instance is better than Windows Azure Macro, medium and high instance while Read, Random read, Write and Random write performance of Windows Azure mini instance is better than OpenStack small instance.

TABLE VIII. OPENSTACK DISK PERFORMANCE BY INSTANCE TYPE

OpenStack (MB/s)				
	Read	Random Read	Write (Random Write
m1.tiny	1232.9069	906.4013386	499.1360657	565.56233
m1.small	1121.3244	776.882588	482.415376	450.896899
m1.medium	1161.6355	912.997511	476.58721	552.165469
m1.large	1147.2248	857.01604	398.637235	542.050332

TABLE IX. WINAZURE DISK PERFORMANCE BY INSTANCE TYPE

Windows Azure (MB/s)				
	Read	Random Read	Write	Random Write
Micro VM	1114.5448	853.210084	410.188796	464.690696
Mini VM	1140.5033	801.040145	527.86001	476.136834
Medium VM	1056.6933	867.744188	418.355458	463.863948
High VM	1082.3363	775.825894	338.068539	486.090585

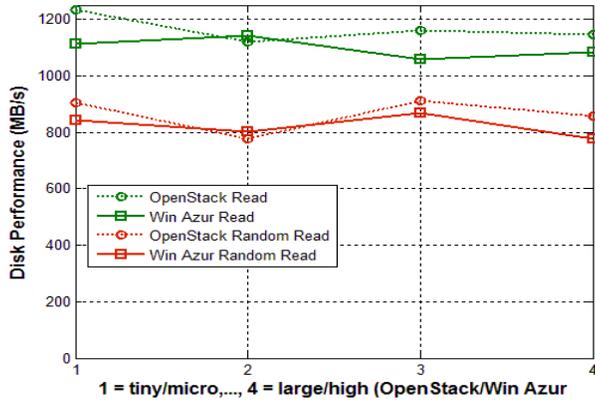


Fig. 6. Disk Read Performance Comparison.

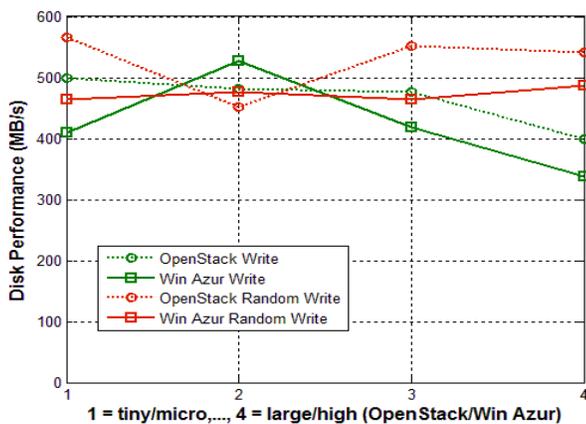


Fig. 7. Disk Write Performance Comparison.

This data provides several facts received from IOzone test. The benchmark provides extremely high speeds information. We notice few causes about disk performance. In virtual machine we have not control over several caches. It might be fine that caching inside the disks or other buffers change these figures. If this happening, then there is a needs to run the throughput experiment once again using a file size that is greater than the amount of free memory. That will effect on the time and these tests may take more than reasonable time.

The primary hardware consumes RAM drive as a buffer to the local hard drive which can make common sense if the actual hard drive is not located jointly with the other hardware holding the application. This is the reality that memory is very costly than hard disk. But alternatively if the primary hardware reports free memory space it uses as disk cache between other things. The random operations are approximately similarly fast as the sequential operations.

The change among the same tasks on different instance hosted by the same cloud provider is vast. We can simply describe it by conflict of resources among tenants unspecified that disk operations are particularly liable to suffer from. Finally, we observe that the systems design appears to be relatively similar between OpenStack and Windows Azure.

D. Network Performance

In network performance test, we take single instances on each platform. UDP was used in order to gather information on packet loss, TCP was used to measure bandwidth and jitter. For network performance comparison of both clouds, we consider three network parameters, i.e. network bandwidth (it refers to the volume of information that can be transferred over a network in a certain amount of time, typically stated in bps), jitter (difference in the latency on a packet flow among two systems, when some packets take more time as compared to others to transfer), and throughput (the average rate of successful message delivery over a communication channel). The outcomes are shown in Tables X, XI and Fig. 8 to 10. The important fact is that OpenStack network points out greater bandwidth statistics. Microsoft Azure internal network underperforms than OpenStack, though not much low. We run Iperf Server on main machines while run Iperf client on virtual Machine in LAN. We observed that network performance of OpenStack is slightly better than WinAzure. Authors of [37] discussed network simulation resources.

To measure network bandwidth (Fig. 8) and jitter (Fig. 9) start Iperf on server in server mode with these parameters

```
>iperf -s -u -P 0 -i 1 -p 5001 -f m
```

and on client side run this command

```
>iperf -c <Server ip address> -u -b 100m
```

Here, -b selection is used to identify the bandwidth to use. Normally Iperf UDP usage 1Mbps we suggest usage complete offered bandwidth to get an idea.

TABLE X. OPENSTACK NETWORK PERFORMANCE BY INSTANCE TYPE

	Bandwidth (Mbit/s)	Jitter (ms)	Throughput (Mbits/s)
m1.tiny	95.1	0.637	69.8
m1.small	95.2	0.627	72.3
m1.medium	95.2	0.613	74.8
m1.large	95.4	0.608	92.1

TABLE XI. WINAZURE NETWORK PERFORMANCE BY INSTANCE TYPE

Windows Azure			
	Bandwidth (Mbit/s)	Jitter (ms)	Throughput (Mbits/s)
Micro VM	95	0.648	68.4
Mini VM	95.1	0.639	70.2
MediumVM	95.2	0.624	72.8
Hi+gh VM	95.2	0.614	89.4

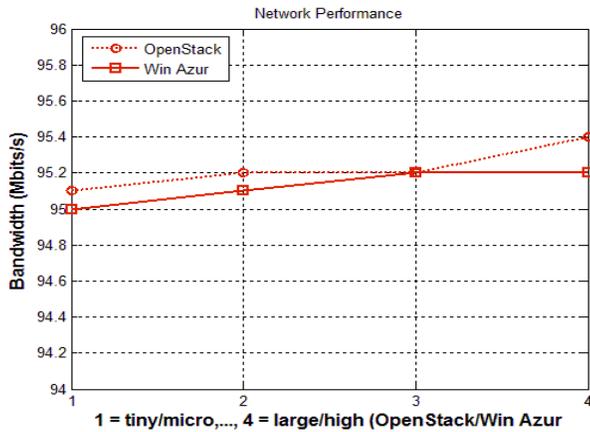


Fig. 8. Network Bandwidth Performance Comparison.

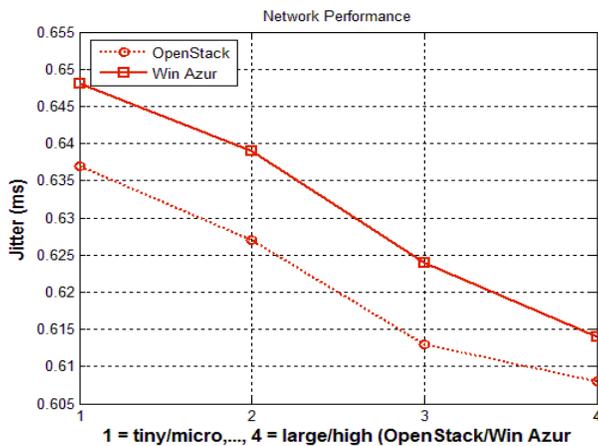


Fig. 9. Network Jitter Performance Comparison.

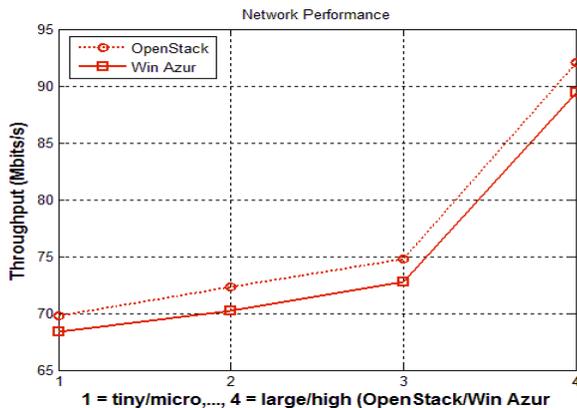


Fig. 10. Network Throughput Performance Comparison.

To measure network throughput (Fig. 10) Iperf on server in server mode with these parameters

```
>iperf -s -P 0 -i 1 -p 5001 -f m
```

on client side run this command

```
>iperf -c <Server ip address> -P 20
```

Where P indicates maximum Parallel TCP connections (20 in this case) to measure throughput of LAN.

In output screen we get 20 different ports on the client is linked to the default 5001 port with the server. Every link has different transfer rate and finally we receive sum of throughput in Mbits/s.

E. Overall Summary of Results

Table XII presents overall picture of performance comparison of two considered cloud platforms w.r.t. CPU, memory, disk and network tested on related benchmarking applications.

Microsoft's Azure platform performs better in terms of CPU-integer Speed on single as well as multi core as compared to OpenStack, whereas in CPU-Floating point regarding single core, though Windows Azure is fractionally better as compared to OpenStack under macro, medium and high instances but its performance is comparatively less under small instance. Considering Floating Point speed with multicore, OpenStack performs a bit better. Overall, Windows Azure in term of CPU performance is marginally better than that of OpenStack when tested through Geekbench benchmark, whereas when checked with LINPACK benchmark, OpenStack's performance is better under tiny and small instances. In Table XII, WA is used for Windows Azur, OS for OpenStack, SC for single core, MC is used for multicore, B for bandwidth and T is used for throughput.

In term of memory performance, Windows Azure performs better under all instances when tested with Stream benchmark. With RAMspeed benchmark, the integer and floating Point memory performance of windows Azure is more under tiny/macro and medium instances while under of small/mini and large/high instances, OpenStack turns out to be better.

With respect to disk Performance in terms of read, random read, write and random write operations, OpenStack is better than Windows Azure under all instances except small one.

OpenStack outperforms Windows Azure in network wise performance in terms of Bandwidth, jitter and throughput.

In summary, the performance of both clouds is almost same but OpenStack is somehow better option keeping in view its cost as well as network performance.

TABLE XII. OVERALL SUMMARY OF RESULTS

Performance Parameters	CPU				Memory (speed in MB/s)			Disk Performance (MB/s)				Network			
Benchmark	Geekbench				LINPACK	RAMspeed		STREAM	IOzone				Iperf		
Instance	Integer		Floating point		GFLOPS (Avg)	Integer Avg	FP Avg	Avg	Read	Random Read	Write	Random write	Bandwidth (Mbs)	Jitter (ms)	Throughput (Mbs)
	SC	MC	SC	MC											
Tiny/Macro	WA	WA	WA	OS	OS	WA	WA	WA	WA	WA	WA	WA	OS	OS	OS
Small/Mini	WA	WA	OS	OS	OS	OS	OS	WA	WA	WA	WA	WA	OS	OS	OS
Medium	WA	WA	WA	WA	WA	WA	WA	WA	OS	OS	OS	OS	OS	OS	OS
Large/High	WA	WA	WA	OS	WA	OS	OS	WA	OS	OS	OS	OS	OS	OS	OS

VI. CONCLUSION

This work explored the performance of two real clouds, namely OpenStack and windows Azure by setting up real time configuration by research team. Performance is one of the major concerns for customers; therefore, we concentrated on observing the performance of major cloud resources like CPU, memory, disk and network using suitable benchmarks for each. These techniques are used to find the issues where the cloud's performance depends on installed software. This research is implemented in a cloud platform where environment load is usually challenging to identify and uncontrollable. The Performance is decreased due to virtual machines hosted on a same physical server running resource-intensive tasks. The performance is decreased in CPU and memory slightly due to intensive task assignment but the difference is more in disk and network intensive task assignment.

We evaluated cloud computing resources from customer's perspectives as well. We tested performance of OpenStack and Microsoft Windows Azure with the help of popular benchmarking tools which are open source and freely available over internet. In some instances, OpenStack performance is more than Windows Azure and vice versa. Overall the performance of Windows Azure and OpenStack cloud is almost same but at some point windows Azure performance is slightly better than OpenStack cloudbut the network performance of OpenStack is much better than Windows Azure. However, in our opinion, OpenStack, being freely available, requiring less infrastructure deployment, less power consumption and incurring less licensing cost is better option.

As a future work, elastic cloud resource management in real cloud environment is good choice for performance analysis of OpenStack and Windows Azure. Authors of [38] provided some hints in this respect.

Virtual machine and network isolation is also very hot issue in the performance of cloud, as a number of machines are using

the same hard drive and network which may impact on performance of a cloud. When many users transfer data these resources are over utilized as a result of performance cloud.

The cloud evaluation is examined on single organization, more research and analysis of varying sizes organizations and in different industries is required to observe its overall application. Further, more research is required on how to find the most suitable cloud solution for a certain organization and system. To support this conclusion more research for the long term effects on organizations of implementation for cloud computing is required.

ACKNOWLEDGMENT

This research is partially funded by Directorate General Monitoring and Evaluation Planning & Development Department, Government of The Punjab, Pakistan, under Government of the Punjab's scheme titled "Capacity Building of Directorate General Monitoring & Evaluation (DG M&E) For Improved Project Planning, Monitoring and Evaluation of Development projects in Punjab".

REFERENCES

- [1] Rajkumar Buyya, Saurabh Kumar Garg, and Rodrigo N. Calheiros, SLA-oriented resource provisioning for cloud computing challenges, architecture, and solutions, proceedings of the 2011 international conference on cloud and service computing pages 1-10.
- [2] Rajkumar Buyya and Karthik Sukumar, platforms for building and deploying applications for cloud computing, csi communications, May 2011, Vol. 35, No. 1. Pp. 6-11.
- [3] Rohit Kamboj, Anoop Arya, Openstack: Open Source Cloud Computing IaaS Platform, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.
- [4] Zach Hill, Jie Li, Ming Mao, Arkaitz Ruiz-Alvarez, and Marty Humphrey, early observations on the performance of Windows Azure, scientific cloud computing. Chicago, illinois, June 21, 2010.
- [5] Vineetha, V. (2012), performance monitoring in cloud, Infosys view point, pp.1-8. <http://www.infosys.com/engineering-services/features/opinions/Documents/cloud-performance-monitoring.pdf>

- [6] Maximo Alves, Luigi Corsello, Daniel Karrenberg and more, new measurement with the RIPE NCC test traffic measurement setup, proc. of PAM 2002, Colorado, USA, 2002.
- [7] T. Lindh, a new approach to performance monitoring in IP networks-combining active and passive methods, proc. of PAM 2002, March, 2002.
- [8] Hong Soon-Hwa, Jae-Young Kim, Bum-Rae Cho and James W. Hong, distributed network traffic monitoring and analysis using load balancing technology. In Proc. of, pp. 172-183. 2001.
- [9] Feifei Chen, John Grundy, Jean-Guy Schneider, Yun Yang and Qiang He, automated analysis of performance and energy consumption for cloud applications, ACM/spec international conference on performance engineering, Dublin, Ireland, Mar 23-26 2014.
- [10] Lefebvre, L. and Orgerie, A.C., designing and evaluating energy efficient cloud, journal of supercomputing, 51(3):352-373, 2010.
- [11] Liu, L., Wang, H., Liu, X., Jin, X., He, W., et al., green cloud: a new architecture for green data center, in proceedings of the 6th international conference industry session on autonomic computing and communications industry session (icac-indst '09), pages 29-38, Barcelona, Spain, 2009.
- [12] Qingwen Chen, Grosso, P.; Van Der Veldt, K.; de Laat, C.; Hofman, R.; Bal, H., profiling energy consumption of VMS for green cloud computing, in proceedings of the 9th IEEE international conference on dependable, autonomic and secure computing(dasc2011), pages 768-775, Sydney, Australia, 2011.
- [13] W. D. Collins, P.J. Rasch, B. A. Boville, J. J. Hack, J. R. McCaa, D. L. Williamson, J. T. Kiehl, B. Briegleb, description of the NCAR community atmosphere model (cam3.0), NCAR report, boulder. 2004.
- [14] Raj Jain, The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. Wiley, 1991.
- [15] R. Buyya, C. S. Yeo, and S. Venugopal, Market- oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities, in proceedings of the 10th IEEE International Conference on High Performance computing and Communications, 2008.
- [16] Paul Brebner and Anna Liu, Performance and Cost Assessment of Cloud Services, Proc of Int. Conference on Service Oriented Computing, 2010, pp. 39-50.
- [17] Q. Zhang, L. Cheng, and R. Boutaba, Cloud computing: state of the art and research challenges, journal of internet services and applications, vol. 1, issue 1, 2010, pp. 7-18.
- [18] Rajkumar Buyyaa, Chee Shin Yeo, Srikumar Venugopal, James Broberg and Ivona Brandic, cloud computing and emerging it platforms: vision, hype, and reality for delivering computing as the 5th utility, future generation computer systems, vol. 25, 2009, pp. 599-616.
- [19] C. N. Hofer and G. Karagiannis, Cloud Computing Services: taxonomy and comparison, journal of internet services and applications (s i: future net service models & designs), vol. 2, 2011, pp. 81-94.
- [20] B. P. Rimal, E. Choi, and I. Lumb, a taxonomy and survey of cloud computing systems, presented at fifth international conference on inc, ims and idc, Seoul, Korea, 2009.
- [21] Simon Ostermann, Alexandria Iosup, Nezh Yigitbasi, Radu Prodan, Thomas Fahringer, Dick Epema, A performance analysis of ec2 cloud computing services for scientific computing, In CloudComp, Volume 34, pages 115-131, 2010.
- [22] Ang Li, Xiaowei Yang, Srikanth Kandula and Ming Zhang, cloudcmp: comparing public cloud providers proceedings of the 10th ACM SIGCOMM conference on internet measurement pages 1-14.
- [23] Li, Zheng, et al., A factor framework for experimental design for performance evaluation of commercial cloud services. arXiv preprint arXiv:1302.2203(2013).
- [24] Ming Mao, Marty Humphrey, a performance study on the VM startup time in the cloud, 5th international conference on cloud computing, IEEE 2012
- [25] Ang Li, Xiaowei Yang, Srikanth Kandula, and Ming Zhang, cloudcmp: comparing public cloud providers, in internet measurement conference, pages 1-14, 2010.
- [26] Y. Lee, A. Zomaya, energy efficient utilization of resources in cloud computing systems, the journal of supercomputing, 60(2) (2012), 268-280.
- [27] Qi Zhang and Lu Cheng, R. Boutaba, cloud computing: state-of-the-art and research challenges, journal of internet services and applications, 1(1) (2010), 7-18.
- [28] Radu Tudoran, Alexandru Costan, Gabriel Antoniu, Luc Bouge. A performance evaluation of azure and nimbus clouds for scientific applications, proceedings of the 2nd international workshop on cloud computing platforms, ACM New York, USA, article no. 4, 2012.
- [29] Ang Li, Xiaowei Yang, Srikanth Kandula and Ming Zhang, cloudcmp: comparing public cloud providers proceedings of the 10th ACM SIGCOMM conference on internet measurement pages 1-14.
- [30] K.Kostantos, A. Kapsalis, D.Kyriazis, M.Themistocleous and P. R. Cunha, open-source IaaS fit for purpose: a comparison between opennebula and openstack, international journal of electronic business management, vol. 11, No. 3, pp. 191-201 (2013).
- [31] S. Itnal, V. Maan, S. Jhajharia, D. Joshi, network performance analysis and optimization on cloud, in proceedings of IRF international conference, 13th April-2014, Pune, India.
- [32] G.v. Laszewski, J. Diaz, F. Wang, Geoffrey C. Fox, comparison of multiple cloud frameworks, cloud computing IEEE 5th international conference on 24-29 June 2012 page734 - 741.
- [33] R. Ledyayev and H. Richter, high performance computing in a cloud using OpenStack, cloud computing 2014: the fifth international conference on cloud computing, Grids, and virtualization.
- [34] Ajiro, Y., & Tanaka, A. (2007, December). Improving packing algorithms for server consolidation. In *Int. CMG Conference* (Vol. 253, pp. 399-406).
- [35] Mehmood, M. A., Mahmood, A., Khan, M. N. A., & Khatoun, S. (2016). A scenario-based distributed testing model for software applications. *International Journal of Advanced and Applied Sciences*, 3(10), 64-71.
- [36] Singh, S., Singh, N., & Singh, S.B. (2016). On a hybrid practical swarm optimization algorithm. *International Journal of Advanced and Applied Sciences*, 3(12), 96-105.
- [37] Esparcia, J. A. P., & Singh, M. (2017). Comprehensive study of multi-resource cloud simulation tools. *International Journal of Advanced and Applied Sciences*, 4(7), 29-38.
- [38] Ngenzi, A., Selvarani, R., & Suchithra, R. (2016). Improvig server consolidation using physcal consolidation concept. *International Journal of Advanced and Applied Sciences*, 3(6), 95-99.

Intelligent Model Conception Proposal for Adaptive Hypermedia Systems

Mehdi TMIMI

Laboratory of Computing and
Interdisciplinary Physics, ENS,
USMBA
Fez, Morocco

Mohamed BENSLIMANE

Transmission and Treatment of
Information Laboratory, EST,
USMBA
Fez, Morocco

Mohammed BERRADA, Kamar

OUAZZANI

Laboratory of Computing and
Interdisciplinary Physics, ENS,
USMBA
Fez, Morocco

Abstract—The context of this article is to study and propose solutions for the major problems of adaptive hypermedia systems. In fact, the works and models proposed for these systems are made according to the tradition of studying first theories and rules, then modeling and designing a system that implements them. As a result, adaptive hypermedia systems designed reflect and support only the elements and information that were studied during the design phase. Also, these systems require a huge amount of data to power their architecture in order to start operating. This famous problem is called “cold start” and until now represents a challenge. So, in this paper, we will propose an intelligent and flexible model inspired by human nature and that proposes a promising solution to these problems concerning hypermedia adaptive systems.

Keywords—Adaptive hypermedia; artificial intelligence; deep learning; learner model; domain model; adaptation model; brain; neuron

I. INTRODUCTION

Adaptive systems have been of great importance in today's world, and are constantly improving by using advances in Internet technologies, psychology, artificial intelligence and education [1].

These adaptive systems can be defined as a learning management system (LMS) or even an individual learning platform that adapts the teaching to individual learner differences, such as cognitive abilities, learning styles, emotional states, etc.

The adaptive hypermedia is one of the adaptive learning systems that has emerged in response to the earlier traditional learning systems based on the "one design for all" [2] and "just on the web" approach. Its main characteristic is to give learners an active role [3] in building their knowledge and skills by adapting all the structural, visual and contextual aspect of the learning process.

To achieve these ends, three interrelated models are required:

1) *Learner model*: describes learner characteristics such as: skills, knowledge, learning styles, etc. [4]

2) *Adaptation model*: describes the set of construction and

presentation rules that are responsible for constructing the content to be delivered to the learner.

3) *Domain Model*: describes how the information content of the application is structured [5].

These three models mentioned above represent the core, the resources and the fundamental elements for the functioning of the adaptive hypermedia, these elements are first collected from scientific and academic works then connected and structured in a logical conception that we call a model.

After studying and proposing new conceptions for each one of the three models, which we published in several papers [6] [7], we intended to start the development of our hypermedia adaptive. However, we first had to deal with two major problems, which are: conceptual flexibility and cold start.

As the days go by, scientific research brings us new elements and relationships that influence the learning process and that need to be added manually and then applied by our adaptive hypermedia system. Unfortunately, and as known, such frequent conceptual changes are often costly in terms of performance and time and risky in terms of data integrity.

Also, the works and references that form the basis of our models may not be fully applicable and depend on the context.

Concerning the "cold start" problem, the main challenge has always been how to collect data about learners without disturbing them with questionnaires, especially in the early interactions with the system when learners can lose motivation and get bored if we welcome them with pre-tests and quizzes.

To overcome these problems, we drew inspiration from human nature (the learner) itself. And we propose for adaptive hypermedia a scalable and flexible system that mimics human complexity with the ability to learn and make decisions automatically.

II. INSPIRATION AND PROJECTION

At the early stage of our reflection, we thought of a smart model that learns and automatically makes decisions. In fact, that kind of model looks similar to the work done on artificial intelligence and deep learning. But we did not take that path and we just observed ourselves and thought deeply about human nature, which is very complex in itself.

A. Inspiration

We; as humans; all our memories, experiences, abilities, and knowledge are managed mainly by our brain.

The human brain is very organized; it is composed of several parts that each has specific roles while being complementary to each other.

This complementarity is done by neurons. In fact the total number of neurons was estimated at 85 billion [8], but recent studies [9] estimated a total of 105 billion neurons in the human cerebellum alone [10].

In Fig. 1, we show the internal composition of a neuron

and how the communication between different neurons is done. In fact, each neuron has a cell body, dendrites and an axon. The communication between two neurons is done by chemical and electrical processes, the electrical nerve messages that arrive at the end of the nerves are found in a neuron, and trigger a secretion of chemical molecules that will deposit on the next neuron, where they will be translated again into an electrical message in the other neuron [11].

Finally, neurons have the ability to learn and depending on the type of learning, the connections between the neurons involved are either “significant or stronger” or “less and weaker” (as shown in Fig. 2).

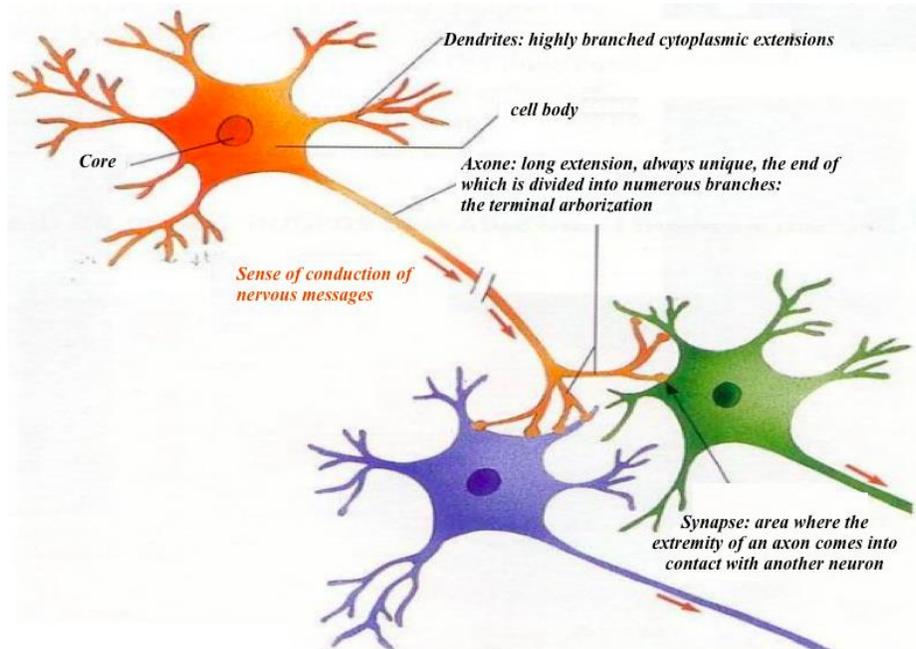


Fig. 1. The Different Elements of a Neuron and how the Communication is Done [11].

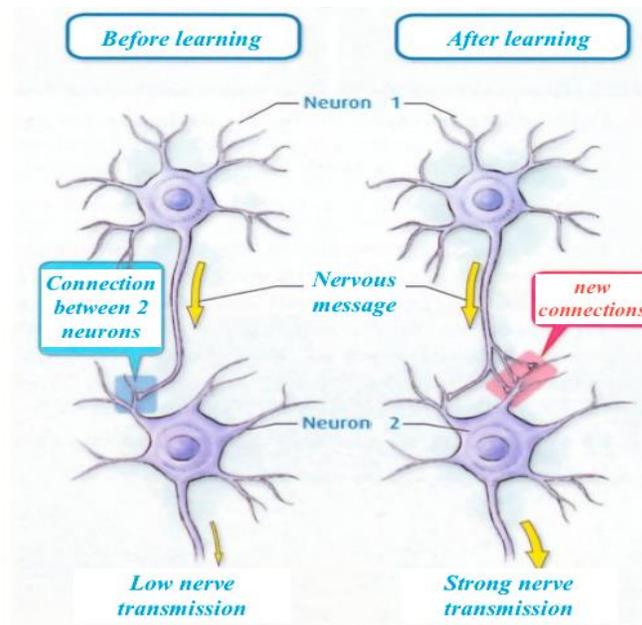


Fig. 2. Neurons before and after Learning [12].

B. Projection

The above was just some definitions and explanations of how the human brain works, let's now do a Reformulation of this complexity while making a projection on our field of research.

Since each neuron represents an information, we will follow the same logic considering that any element of our system will play the role of a neuron. This element refers to concrete and unique information within our system and will be defined as a couplet (element name, element value).

Concerning the communication between the neurons, we will implement it as logical and reflexive relationships between the elements, where they can play both the role of inputs (like neuron dendrites) or outputs (like neuron axons) for these relations.

Finally, and as we mentioned earlier concerning the learning ability of neurons, we thought of a coefficient mechanism that will indicate the strength of connections and relationships between elements.

III. PROPOSITION

In the following and as shown in Fig. 3, we will present our intelligent model based on the projection discussed above and designed with UML2 class diagram.

We will first present the different components and relationships of our model, then we will explain the constraints and rules necessary for its operation and finally we will detail

how to implement them while presenting the main use case and a future extension of our model once it is in action.

A. Components and Relations

Our model is composed of three classes, which are:

- 'Element': this class represents an information within our system and it is characterized by a unique identifier, name, value and a description. The identifier [id property] will uniquely reference this information. Also, each instance of the element class can simultaneously play many 'input' and 'output' roles in different relations.
- 'Relation': This class is characterized by a unique identifier, name, rule and a date. The rule property defines the actions and operations to be performed for all the results of this relation, the allowed values are: {Update, Delete and no Action}. Also, each instance of the Relation class can have one or more Result, and an instance of result class concerns one and only one Relation.
- 'Result': this class is characterized by a unique identifier, coefficient, state and a date. The coefficient property; which is an integer value; varies from 1 to ∞ depending on the number of times that the system has approved this result. Finally the state property; which is a string value; describes the state of this relationship. The allowed values are: {current, historic}.

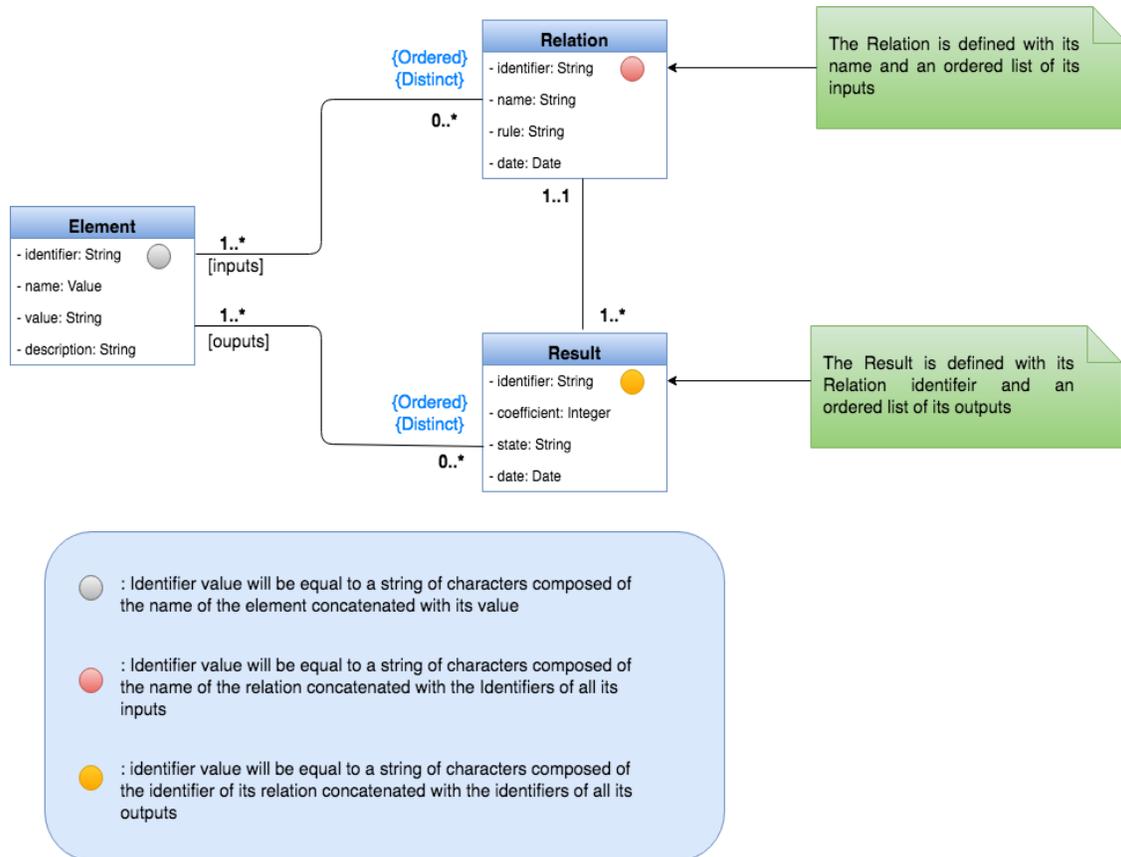


Fig. 3. Class Diagram of Our Proposed System.

B. Constraints and Rules

All the constraints and rules that we will explain are developed to ensure the uniqueness of some components of our model.

- The first rule concerns the uniqueness of the element, i.e. it is impossible to have two elements with the same identifier.
- The second rule concerns the uniqueness of a relation, which is defined by its name and an ordered and distinct list of its inputs. This rule will make it impossible and forbidden to have two relations that have the same names and the same Inputs. In Fig. 4, we present a concrete explanation of this rule where we have a 'Relation: {r1}', which has "define" as a name and two inputs elements e1 and e2. By involving the constraint cited above the 'Relation: {r4}' can not be created and the system will consider it as a 'Relation: {r1}'
- The third rule concerns the uniqueness of a Result, which is defined by its relation name and an ordered and distinct list of its outputs. This rule will make it impossible and forbidden to have two Results that have the same Relation and the same Inputs. In Fig. 5, we present a concrete explanation of this rule; Whenever the system approves an existing Result, it will be translated by incrementing the coefficient of the 'Result: {res1}' instead of creating a new 'Result: {res3}'.

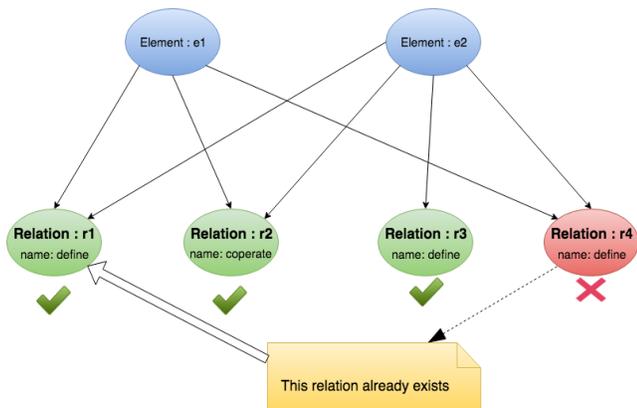


Fig. 4. Diagram Expressing the Constraint of the Relation Identifier.

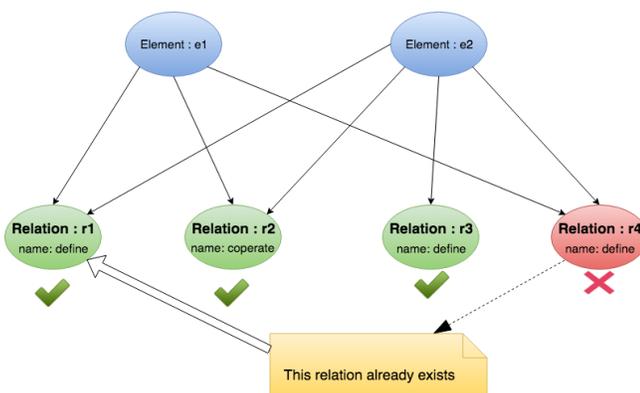


Fig. 5. Diagram Expressing the Constraint of the Result Identifier.

C. Implementation Rules

The major difficulty in expressing the rules explained above is summarized in how to ensure the uniqueness of an element, a relation and a result?

In fact, we used the relational data logic model (MLD-R) for this purpose and we worked on algorithms that will ensure the generation of the primary keys.

In the following, we will present for each rule the mechanism of generation of its identifier.

- Concerning the uniqueness of the element, the identifier will equal to a string of characters composed of the element name concatenated with its value. **Element Identifier** = Element name + Element value
- Regarding the uniqueness of the relation, the identifier will equal to a string of characters composed of the Relation name concatenated with a sorted and distinct sequence of the input element identifiers.

$$\text{Relation Identifier} = \text{Relation name} + \text{Element1 Identifier} + \text{Element2 Identifier} + \dots + \text{ElementN Identifier}$$

We added {ordered} and {distinct} constraints to the list of input element in the class diagram (Fig. 3) to ensure that we will not have in any case, two Relations that have the same name and the same list of input element and also to prevent the same element to play multiple input roles for the same relation.

Also, at the moment when the system approves or triggers a Relation, our system will first generate the identifier dedicated to this Relation. Then, from this generated identifier, the system will look for this relation, and following the result of this query our system will decide either to create a new relation or just use the existing one.

- Concerning the uniqueness of the Result, the identifier will equal to a string of characters composed of the Relation identifier concatenated with a sorted and distinct sequence of the outputs elements identifiers.

$$\text{Result Identifier} = \text{Relation Identifier} + \text{Element1 Identifier} + \text{Element2 Identifier} + \dots + \text{ElementN Identifier}$$

We added {ordered} and {distinct} constraints to the list of output element in the class diagram (Fig. 3) for the same purpose explained above.

Also, when the system approves or triggers a Result, our system will first generate the identifier dedicated to this Result. Then from this generated identifier, the system will check if we already have a Result with this identifier, and following this query our system will either decide to create a new Result and initialize its coefficient to 1 or just use the existing Result while incrementing its coefficient.

D. Main Use Case and Future Extension

Before discussing the main use case, we will first show the location of our intelligent model within the global architecture of adaptive hypermedia.

In order to do that, we show in Fig. 6 the global architecture proposed by the Munich reference model [13],

which consists mainly of three models (user / learner, domain and adaptation model).

Unlike this architecture proposed by Munich, we will introduce our intelligent model while keeping the same architecture composed of the three models.

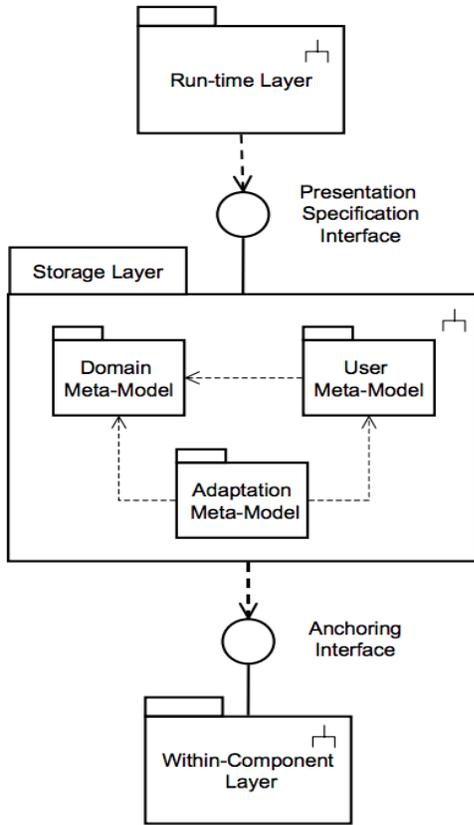


Fig. 6. Munich's Architecture of Adaptive Hypermedia Applications [13].

In fact, we have planned two phases for our system:

The first is the learning phase of our intelligent model, during this phase the adaptive hypermedia is based primarily on the three models while our intelligent model is simply observing and learning by defining and connecting the different elements of these three models (Fig. 7). In fact, we imagine our proposed model as a baby that needs to be educated first, and over time it will begin to apply what it has learned.

The second phase represents the final phase where our intelligent model takes a step forward and becomes the main model that drives the adaptive hypermedia while the three models become secondary (Fig. 7) and considered as its support.

Moreover, the mechanism of the coefficients that we have implemented in our system was made to imitate the consciousness and the subconsciousness of the human being.

So, whenever we force our system to trigger a result or supply it with data, the system is considered to be in a consciousness mode while the subconsciousness mode is reserved for autonomous decisions and behaviour made by our system.

Regarding the problem of cold start, we thought from the beginning to reorient the main challenge of "how to collect data on learners without disturbing them with questionnaires?" to "how to make decisions using the minimum of data?"

Indeed, from all the previous explanations, we can clearly see that our system forms a network of connected elements that continues to grow while having full flexibility in terms of structure and depth.

As a result, the bigger it grows, the less input data is required to make decisions.

Finally, we thought of a future extension of our model. This extension will be based on our current model to generate a network of abstract elements. By abstract element we refer to the definition or Meta data of the element.

This extension will present all our global rules and theories established which are based on the experiences learned by our system.

IV. CONCLUSION AND PRESPECTIVES

Our intelligent system will operate side-by-side with existing models within the hypermedia adaptive systems while enabling existing ones to become more dynamic and flexible by overcoming the two main conceptual flexibility and cold-start issues.

Also, the extension of the scope of our intelligent model is still at the theoretical phase, therefore we intend to work on its implementation in order to have a unique and independent model that can replace all other models on which adaptive hypermedia is based.

Finally, we plan in our next work to develop the global architecture of our adaptive hypermedia by linking all our proposed published models (learner model, adaptation model,

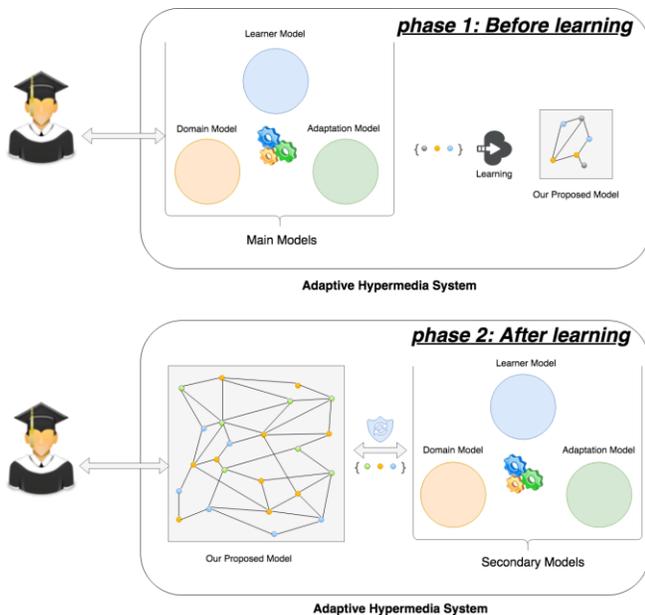


Fig. 7. Explanatory Diagram of the Two Phases of Our Intelligent Model within the Adaptive Hypermedia.

domain model and our proposed intelligent model) and then develop and test our system, which is the final goal of our research.

REFERENCES

- [1] Kara, N., & Sevim, N. (2013). Adaptive Learning Systems: Beyond Teaching Machines. *Contemporary Educational Technology*, 4(2), 108–120.
- [2] G. Kaya and A. Altun, "A Learner Model for Learning Object Based Personalized Learning Environments," *Commun. Comput. Inf. Sci.*, vol. 240, pp. 349–355, 2011.
- [3] A. Behaz, M. Djoudi, "Approche de Modélisation d'un Apprenant à base d'Ontologie pour un Hypermédia adaptatif Pédagogique", In: CIIA, 2009.
- [4] D. Samia, B. Taher, and L. Yacine, "Une nouvelle approche pour l'adaptation d'un hypermédia pédagogique au profil cognitif de l'apprenant en utilisant XML," in *2nd Conférence Internationale sur l'Informatique et ses Applications (CIIA'09)*, 2009.
- [5] P. M. E. Wu, H.; Houben, G.J.P.M.; De Bra, "Supporting user adaptation in adaptive hypermedia applications," in *P. Vet, van der, & P. M. E. De Bra (Eds.), Proceedings Conferentie Informatiewetenschap 2000*, 2000, vol. Vol. 00-20, pp. 88–98.
- [6] Tmimi, M., Benslimane, M., Berrada, M., & Ouazzani, K. (2017). A Proposed Conception of the Learner Model for Adaptive Hypermedia. *International Journal of Applied Engineering Research*, 12(24), 16008–16016.
- [7] Tmimi, M., Benslimane, M., Berrada, M., & Ouazzani, K. (2018). Implemented and Tested Conception Proposal of Adaptation Model for Adaptive Hypermedia. Accepted paper for publication in the *International Journal of Emerging Technologies in Learning (IJET)*
- [8] Williams RW, Herrup K (1988) *The control of neuron number*. *Annu Rev Neurosci* 11:423–453.
- [9] Andersen BB, Korbo L, Pakkenberg B (1992) A quantitative study of the human cerebellum with unbiased stereological techniques. *J Comp Neurol* 326:549–560
- [10] Suzana Herculano-Houzel ,Roberto Lent : Isotropic Fractionator: A Simple, Rapid Method for the Quantification of Total Cell and Neuron Numbers in the Brain, *The Journal of Neuroscience*, March 9, 2005 • 25(10):2518–2521
- [11] *La 3D: Cerveau*, online in: "<http://tpela3d.e-monsite.com/pages/i-la-vision-de-l-homme/cerveau.html>".
- [12] Editeur Belin, *Svt, Iere s*, Manuel de l'élève (édition 2011) DUCO, ANDRE, 2011.
- [13] Koch, N., & Wirsing, M. (2002). The Munich Reference Model for Adaptive Hypermedia Applications An Overview of the Reference Model. In *2nd International Conference on Adaptive Hypermedia and Adaptive Web-based Systems* (pp. 213–222).

Cost Aware Resource Selection in IaaS Clouds

Uzma Bibi

Shaheed Zulfikar Ali Bhutto Institute of Science and Technology,
Islamabad, Pakistan

Abstract—One of the main challenges in cloud computing is to cope up with the selection of efficient resources in terms of cost. There are various cloud computing service providers which dynamically provide resources to the customers through different pricing policies. Based on the different APIs and pricing policies of the service providers, it becomes difficult for the customers to select the best service provider in terms of cost. In some cases, if the usage of the resources provided by a datacenter exceeds certain limit, then the providers cannot offer more resources to the customers as new VMs cannot be created. Hence, even if the customer chooses the best provider based on the least cost parameter, still there is no guarantee that the provider allocates complete resources to the customer. For this reason, I present system architecture that selects the best service provider based on the customer requirements mainly the cost. The proposed architecture also performs resource management by automatically providing new VMs from the available service providers in the inter cloud. The proposed system is based on five clouds i.e. Amazon EC2, Cloudsigma, Google, GoGrid, and Windows Azure. An interface is designed for obtaining the user requirements. These user requirements are matched with the design database of five cloud providers and based on the matched values; the catalog of optimal costs for each particular cloud is shown to the user. Then Cost Aware Resource Selection algorithm is run for determining the lowest optimal cost for Instance based approach and Quantity based approach. The algorithm tackles two domains of clouds for the algorithm i.e. Single Cloud and Multi Cloud.

Keywords—Cloud computing; pay-as-you-go; infrastructure as a service; cost aware resource selection; virtual machines; hypervisor; instance based approach; quantity based approach; single cloud; multi cloud

I. INTRODUCTION

A. Overview of Cloud Computing

Cloud computing is an emerging paradigm which provides the computing services as fifth utility [1]. Cloud computing provides the resources on pay as you go basis which gives new perspective and identity to the cloud business.

Clients and system administrators can deploy their applications and web services through the allocation of resources by cloud providers. They do not have to invest an upfront cost on Infrastructures and do not have to manage the resources on their own. The tasks are reduced and resources are all managed by the cloud providers providing the resources. The definition of cloud computing was represented by Vaquero [2] after reviewing many different other definitions of cloud computing:

“Clouds are a large pool of easily usable and accessible virtualized resources such as hardware, development

platforms and/or services. These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized SLAs”.

National Institute of Standards and Technology (NIST) has also presented the definition of Cloud computing that is easily understandable [3]:

“Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models”.

Service on demand, network access, elasticity, pooling of resources, and measured service are the five important characteristics present in the NIST definition of Cloud Computing [3]. There are deployment and service models in the cloud environment. The deployment models are public cloud, private cloud, community cloud, and hybrid cloud (the combination of both private and public clouds). The service models are SaaS (Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service). Fig. 1 shows the service models and the deployment models in cloud computing. It summarizes the operation of three service models on the top of deployment models.

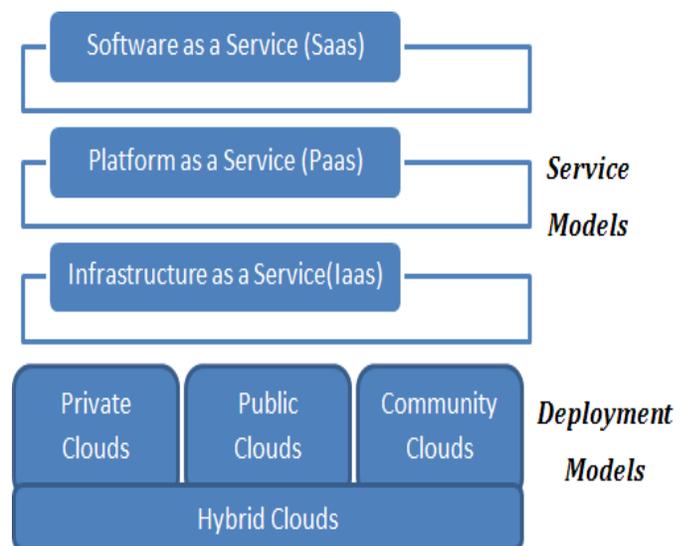


Fig. 1. Service and Deployment Models in Cloud Computing (Source: [4]).

Cloud computing brings some features that are attractive to clients and can change the procedure of accessing the resources from clouds making it easy and convenient. The best feature among all features that cloud computing provides is usage of a service as a utility that is pay as you go model of accessing and using the resources, elasticity in resource provisioning, resource management, and having the infrastructure without investing on it. Broad network access is another feature of cloud computing that lets the clients access the resources from Internet [3, 5].

Section II provides the related work of other authors highlighting their limitations and strengths of my proposed work. Section III discusses the proposed framework and the underlying design. Section IV provides implementation and the experimentation results. Section V draws my conclusions.

B. Cost Aware Decision System for Resource Selection in Clouds

Earlier, there have been some systems presented for the system administrators to select resources and deploy their web applications and services on clouds. Decision system for resource selection [6] is one of such approaches. Through this approach the system administrators have an ease in selecting the best resource provider from all over the world.

The cost aware resource selection decision system maintains a database of various services provided by different cloud providers. We take five cloud providers for our system on general basis; these are Amazon EC2, Google, GoGrid, CloudSigma, and Windows Azure. The system obtains requirements from the clients and then a comparison is done for the various services saved in the database. Based on the services that match the requirements and charge the lowest cost are selected as the pertinent cloud provider for the client. The best resource can be selected from two or more than two cloud providers. Then the computation cost among the modules is evaluated and a comparison is done to suggest the best resource to the users.

The cost aware resource selection decision system is a solution to the problems that a system administrator faces during the resource selection that best matches their requirements. It is assumed that the resources of the applications that are needed to deploy them are already known to the system administrator. This research covers the main important factor i.e. cost. It provides the cost awareness for resource selection in clouds but it does not cover other parameters like performance and is left for further extension to this research.

II. RELATED WORK

Zhang [7] worked on providing the cloud resources on user demands considering the cost factor as minimum as possible. The focus of the research was providing the network, storage, and computing resources. The author managed to work on three layers to provide the best solution. Although the approach was good in providing the cheapest resource provider, it did not provide the complete solution of getting the resources after having the knowledge of the best provider. The users had to manage the resource selection and retrieval at their own. Prashant and Upendra [8] propose a cost aware

system that optimizes the selection of virtual server that minimizes the cost. The researchers also focused on the mechanisms to reduce the time in new configurations. Their main focus was to remove the tradeoff between cost and time for cloud resources in return ignoring the solution to provide the best resource provider offering minimum cost. My proposed system is complementary to this and provides the best cost awareness for efficient resource selection in inter clouds.

Liu and Zhang [9] presented a framework that chooses the resource provider based on the quality of service. The work comprised of discovery agents that managed the discovery of cloud resources using hash indexing. The whole network was divided into different domains to collect the Quality of Service information to choose the best resource provider. The focus was purely abstract covering the quality with the least focus on cost. My system removes the abstraction level, with a complete focus on minimizing the cost.

A survey of cloud computing presented in [10, 11] highlight the architecture and key principles about the resource selection with high performance and minimum cost. These survey papers presented the service selection algorithm that is adaptive to different kind of environments with scalability and availability. The research is only limited to design and architecture of cloud computing for resource selection covering performance and cost factors at a very abstract level. Due to insufficient resource available at times, the cloud provider may not be able to provide the demanded resources to the clients [12]. High cost of resources to meet the requested services by clients has led to the pay-as-you-go model to avoid the fluctuations in cloud computing. Seagull [12] was introduced as a solution to overcome the outburst in cloud computing due to scarcity of resources. This technique focused well on overcoming the sudden outburst but did not provide a unified solution to provide cost awareness to the clients for selecting resources.

Different cloud providers need to work on how to price the cloud resources requested by the clients from different environments [11, 13]. Enhanced ant colony system was introduced by Wang [13] which discussed the composition cost of data transfer for composing a service. A service selection algorithm was introduced which was able to deal with parallel relations between services. This algorithm only focused on the location of service providers and users, neglecting the other parameters which have a great impact on the cost evaluation. My system covers all the basic parameters to evaluate cost of the resources in an efficient way.

Inter cloud architectures and brokering systems for cost efficiency proposed by Nikoley and Rajkumar [14] discuss the inter cloud environments that facilitate the brokering of cloud resources in clouds. Kingfisher [15] is a model that presented the cost aware elasticity of cloud resources. This system focused on reducing the cost of virtual servers and their configurations. Kingfisher removed the tradeoff between cost and time but did not provide the efficient solution to the users to select the best service based on minimum cost. My system presents the unified solution to this problem. Mistral [16] addressed the tradeoff between three parameters i.e. power,

performance, and cost. It is a framework that optimizes these three parameters to maximize the utility. Its focus is on addressing the tradeoff neglecting the parameter cost, which is a major factor in resource selection. Mansoor and Lakshman [17] proposed a resource allocation algorithm for distributed clouds. The objective was to allocate the resource by minimizing the distance between cloud providers and clients, optimizing the selection of servers in clouds. The performance of the algorithm was evaluated through simulations. As the proposed system was based on resource allocation, it did not provide a solution to the cost aware selection of resources. Whereas, my proposed system satisfies the solution to cost aware resource selection in inter clouds.

Meikang [18] proposed a two dynamically based resource allocation algorithms that adjust the updated task information for execution. These algorithms allocate the resources to the clients and schedule the order of execution for tasks. The objective of these algorithms was to improve the performance. OPTIMIS [19] was proposed for dynamic provisioning of cloud resources. The objective of the technique was to optimize the life cycle of a service including its construction, deployment, and operation on the basis of cost and other factors. It aimed at providing a reliable and trustful cloud computing environment. Both the above techniques aimed at enhancing the performance and optimizing the resource provisioning in the cloud, but somehow neglected the cost awareness for resources.

Cloud resources and data centers have been aiming at reducing energy consumption [20]. The technique aims at providing a real time service for a virtual request. Different schemes have been proposed to reduce energy consumptions and enhance the performance through simulations [16, 20]. Dynamic load distribution policies have been proposed by Kien and Jingru [21] that addresses the cost related to electricity and the cooling effects. Load spikes have been handled through different cooling strategies, a comparison have been done between the cost aware and cost unaware policies by addressing the cost saving related to electricity and load migration.

Reliability Profit Assurance (RPA) algorithm [22] was presented to investigate the reliability of resources in distributed computing systems in context of operating costs and scheduling. To increase reliability, RPA algorithm introduced cost aware replication mechanism. This research made a first attempt to evaluate the reliability cost in workflows but made many gaps in the research which are addressed in our system. Resource selection strategy [23] is presented to select a host based on the minimum network delay. The aim of this model is to minimize the time between requesting and retrieving a service from cloud provider by selecting the host that is the closest in the network. And it minimizes the execution time of the tasks. The approach was good but it only focused on the location of the cloud provider, minimizing the distance between the client and provider to decrease the cost whereas our proposed system covers other necessary factors other than location for the best selection of cloud provider that charges the lowest cost.

Resource scheduling and optimization is an emerging

paradigm in the field of cloud computing [24, 25]. The scheduling algorithm has been introduced in the hybrid clouds presenting the important aspects when scheduling workflows [24]. The communication channels are accessed on job allocation and the performance of scheduling algorithms has been evaluated by comparing the impact of the available bandwidth. For scheduling purpose, the tasks are divided and grouped into the requested resource and data, and then are prioritized [25]. Resource selection is done using a priority formula over sequential scheduling but it does not provide the best resource provider that provides cost aware resources on user demands. Cost aware allocation of cloud resources have been presented by Markus and Simon [26]. In their work, workload forecasting model has been introduced based on Fourier transformation. The aim of the paper was to achieve efficiency of resource utilization. The work aimed at scheduling and improving the resource utilization but did not focus on the cost awareness of resource selection in inter clouds, which is best presented in my proposed system. HCOC [28] addressed which resources should be requested from the public cloud and merged with the private cloud to provide the best results within a given execution time. The aim was to achieve the desired result at a given execution time. The results were then evaluated through simulations. A delay-constrained optimization framework [27] has been proposed on the basis of cost models, to minimize the cloud overhead and maximize the resource utilization. The total execution time is reduced by selecting the appropriate mapping nodes for the modules that have been assigned priority.

Selection and binding of resources at an optimal cost is discussed [30]. The focus of the paper was to improve the quality of results by achieving efficiency, robustness, and scalability. A prototype implementation was also presented for the resource selection and binding component. Different parameters were analyzed to estimate the global energy needed for the development of a service or system. The types of energy like operating energy and embodied energy were discussed [29]. But the only factor in focus was -energy- whereas my proposed system focuses on -cost- as the main factor for service selection from anywhere in the world.

Hybrid modeling [31] focuses on the complex structure of resource selection systems having the ability to cope up with the enterprise architecture. It focuses on the simulation techniques to decrease the involvement of stakeholders and third parties as compared to other techniques. But the paper only focused on the scheduling of the services ignoring the factor of cost which is deeply covered in my proposed system.

III. PROPOSED FRAMEWORK

In my proposed solution, I present a new cost aware resource selection algorithm. The user will give the input based on two approaches provided by the algorithm i.e. Instance based and Quantity based approaches. Based on these inputs, the algorithm will display the best cloud providers that charge the lowest cost for the specified services.

A. Proposed Cost Aware Resource Selection Algorithm

The system mainly takes input from the users based on the factors that influence cost. The first part of this research is to

find out those factors and is termed as parameters for which user input is obtained. My system decides the best cloud provider among Amazon EC2, Cloudsigma, Google, GoGrid, and Windows Azure, considering the given user input and displays the optimal costs of every cloud provider. It also displays the lowest cost from the list pointing out the most pertinent cloud provider in terms of cost. The system architecture of the proposed cost aware resource selection system is shown in Fig. 2. The system only focuses on IaaS.

The database designed for the proposed system holds all the detailed information about the cloud providers. It also stores the user requirements in a separate table. The database holds the type of instances of the clouds, the number of cores for every particular instance, RAM (in GB), computational storage (in GB), standard storage, bandwidth in, bandwidth out, location, operating system, contract period and cost for every particular instance and their combinations.

Cloud ecosystem is also integrated with the proposed model. Through cloud ecosystem integration services, cost aware resource selection algorithm can reduce up-front infrastructure capital and maintenance costs while also reducing or keeping their in-house infrastructure footprint or inventory under control.

The users will be enabled to provision Windows and Linux operating systems. The system's consistency will enable them to use the same VMs and management tools on the platform that they use on their premises, thus reducing the costs.

- Virtual machines
- Storage, backup and recovery
- Big compute

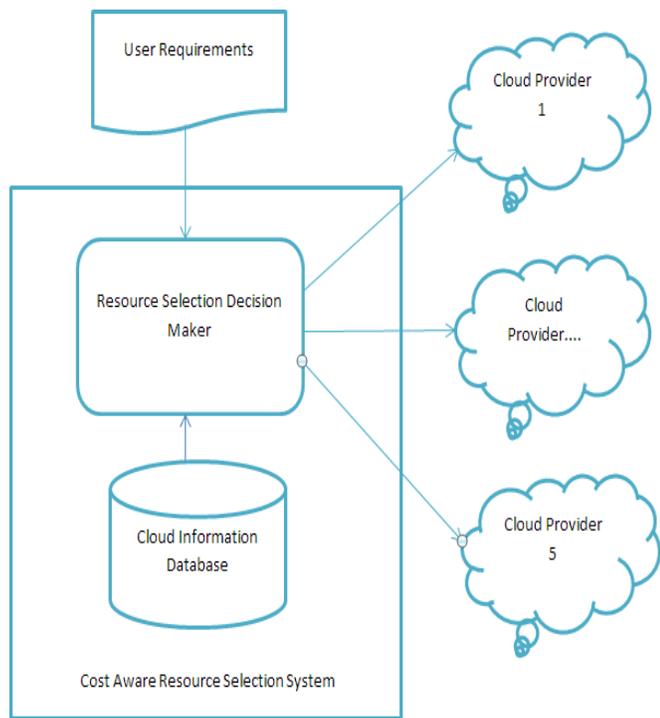


Fig. 2. System Architecture of the Proposed Cost Aware Resource Selection System.

B. Detailed Design

Cost aware resource selection system takes the user requirements based on the identified parameters, analyses them, and provides the most pertinent cloud provider in terms of cost. The components of the system are designed in MVC (Model-View-Controller) model to reduce dependencies among them. The MVC architecture of the system is shown in Fig. 3.

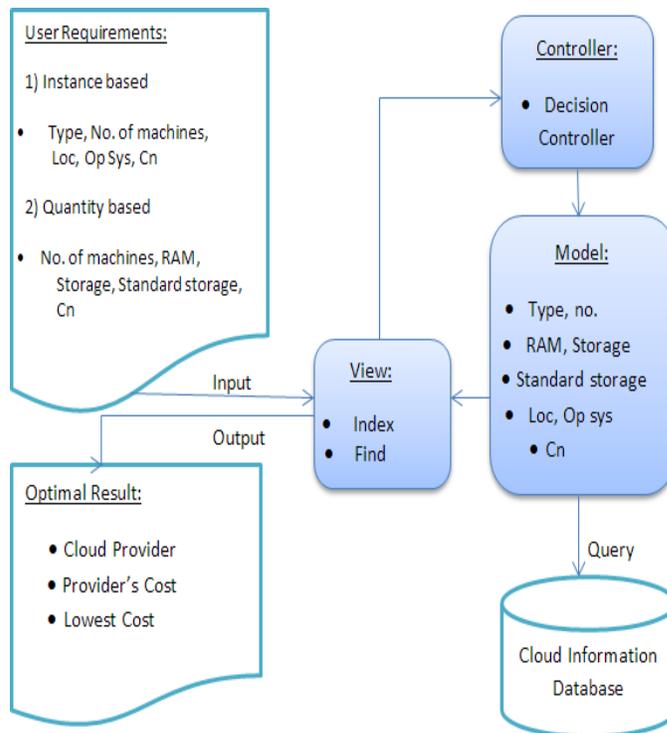


Fig. 3. Cost Aware Resource Selection System Architecture for Model-View- Controller.

TABLE I. FACTORS OF USER REQUIREMENTS

Field	Description	Example
Resource (CPU, RAM, Storage)	Quantity user requires	CPU= 2, RAM= 0.5, Storage= 20
Type of Instances	Every cloud has their particular Instances	t2microinstance
No. of machines	Number of cores user requires	30
Location	User can select US_east or US_west	US_east
Operating System	User can select Linux or windows	Linux
Contract period	It can be hourly, monthly, or yearly	5 hours

IV. IMPLEMENTATION AND EXPERIMENTATION

A. Implementation

Cost Aware Resource Selection system is implemented using Java scripting and PHP. I have implemented this prototype as a proof of the proposed framework. Five cloud providers have been considered for this research work, from the list of IaaS providers, that is Amazon EC2, Cloudsigma, Google, GoGrid, and Microsoft Windows Azure. A web interface is designed for the user to take the requirements as Input. These web pages have been designed for Single cloud user requirements and Mutli-cloud user requirements.

Fig. 4 shows the Single Cloud web interface for the user to enter requirements. This web page has 2 options for the user i.e. Instance based, and Quantity based. If the User selects the Instance based approach, the user is asked to enter the requirements based on Instance type, number of instances, location, operating system, and contract period. The input is taken and is matched with the records of database under every particular cloud through a query. The cost of the record that matches the user requirements is retrieved from the database and then a list is displayed to the user with all the providers and their respected costs. Then a comparison is made between the costs of five cloud providers. The cost which is lowest is displayed to the user.

Fig. 4. Single cloud web interface

In the Quantity based approach section, the user can give the input based on the parameters of number of cores, RAM, Computational storage, contract period. The user can give the standard storage based on the cloud he/she selects. This input is taken and the logic is performed at the run time. In the single cloud, the user can enter the standard storage for a single cloud. The logic is given below:

Amazon Ec2: Storage =
“n” GB Cost per GB= \$ 0.03

So, Storage cost= \$0.03*n

Provider’s Cost= Amazoncost + storage cost

Here Amazoncost is the cost that is retrieved from the database. Similarly, User selects a standard storage of Cloudsigma, Google, GoGrid, and Windows Azure.

Cloudsigma:

Storage =n Cost per

GB= \$ 0.13

So, Storage cost= \$0.13*n

Cost= Sigmacost + storage

cost **Google:** Storage =n

Cost per GB= \$ 0.026

So, Storage cost= \$0.026*n

Cost= Googlecost + storage

cost **GoGrid:** Storage =n

Cost per GB= \$ 0.15

So, Storage cost=

\$0.15*n Cost= Gridcost

+ storage cost **Windows**

Azure: Storage =n Cost

per GB= \$ 0.024

So, Storage cost= \$0.024*n

Cost= Windowscost + storage cost

After taking the input from the user, the list with optimal results is displayed to the user. After displaying a list with five cloud providers and their respective costs, a comparison is done between these costs and the results are shown in Table IV.

B. Algorithm Logic and Flow

Cost aware resource selection algorithm has been split into two basic algorithms. One is through using the Instance based approach while the other is through using Quantity based approach. Using instance-based approach; the user provides the input in the form of type of instance, number of instances, location, operating system, and contract period. The variable “n” defines the number of additions to an instance. The maximum limit has been defined as 10.

Optimal cost is defined as the cost of the output from the cloud provider’s database based on the factors like RAM (GB) CPU cores, computational storage, location, and the communication cost, computation cost, additional storage cost, cost of network bandwidth in and cost of network bandwidth out. Mathematically we can formulate as follows: Let Database cost= DBC, Communication cost= ComC, Computation cost= CtC, additional storage cost= ASC, Network bandwidth In Cost= NBIC, Network Bandwidth Out Cost= NBOC.

So, Optimal Cost= DBC+ComC+Ctc+ASC+NBIC+NBOC

Algorithm 1: Cost aware resource selection algorithm using Instance based approach

Input: maxIns: the number of maximum iterations for instances;

Ti: type of instance;

Ni: number of instances; Li: location;

Osi: operating system; Cni:

contract period;

Output: C1, C2, C3, C4, C5: Costs of Amazon EC2, Cloudsigma, Google, GoGrid, Windows Azure

Lc: Lowest optimal cost

```
1. Lc=NULL; length=0; n=length+1;
2. while (n<10) do
//Td, Ld, Osd are the saved values of particular cloud providers
3. if Ti=Td AND Li=Ld AND Osi=Osd then
3.1. set C1= saved cost*N1*Cn1;
3.2. set C2= saved cost*N2*Cn2;
3.3. set C3= saved cost*N3*Cn3;
3.4. set C4= saved cost*N4*Cn4;
3.5. set C5= saved cost*N5*Cn5;
4. else C1&C2&C3&C4&C5 = NULL;
5. end while
6. if C1≠0 then
6.1. if C1<C2 AND C1<C3 AND C1<C4 AND C1<C5 then
6.2. UPDATE Lc=C1;
6.3. end if;
7. if C2≠0 then
7.1. if C2<C1 AND C2<C3 AND C2<C4 AND C2<C5 then
7.2. UPDATE Lc=C2;
7.3. end if;
8. if C3≠0 then
8.1. if C3<C1 AND C3<C2 AND C3<C4 AND C3<C5 then
8.2. UPDATE Lc=C3;
8.3. end if;
9. if C4≠0 then
9.1. if C4<C1 AND C4<C2 AND C4<C3 AND C4<C5 then
9.2. UPDATE Lc=C4;
9.3. end if;
10. if C5≠0 then
10.1. if C5<C1 AND C5<C2 AND C5<C3 AND C5<C4 then
10.2. UPDATE Lc=C5;
10.3. end if;
11. end if;
12. return Lc;
```

Now, the second algorithm states the Quantity based approach. The user gives the input inform of quantities like number of machines, the amount of RAM, computational storage, the standard storage which is unique for every particular cloud and for that the user chooses the cloud too, and finally the contract period.

Algorithm 2: Cost aware resource selection algorithm using Quantity based approach

Input: Ni: number of machines; Ri: RAM; Si: computational storage; Ssi:

standard storage;

Ci: cloud

provider Cni:

contract

period;

Output: C1, C2, C3, C4, C5: Costs of Amazon EC2, Cloudsigma, Google, GoGrid, Windows Azure

Lc: Lowest optimal cost

```
1. Lc=NULL; C1=0; C2=0; C3=0; C4=0; C5=0;
2. while input fields are not empty do
//Td, Ld, Osd are the saved values of particular cloud providers
3. if Ri=Rd AND Si=Sd then
3.1. if Ci= Amazon EC2 then;
3.1.1. set A=0.03*SSi;
3.1.2. UPDATE C1= saved cost*Cni*Ni + A;
3.2. if Ci= Cloudsigma then;
3.2.1. set C2=0.13*SSi;
3.2.2. UPDATE C2= saved cost*Cni*Ni + B;
3.3. if Ci= Google then;
3.3.1. set C=0.135*SSi;
3.3.2. UPDATE C3= saved cost*Cni*Ni + C;
3.4. if Ci= GoGrid then;
3.4.1. set D=0.15*SSi;
3.4.2. UPDATE C4= saved cost*Cni*Ni + D;
3.5. if Ci= Windows Azure then;
3.5.1. set E=0.024*SSi;
3.5.2. UPDATE C5= saved cost*Cni*Ni + E;
4. else C1&C2&C3&C4&C5 = NULL;
5. end while
6. if C1≠0 then
6.1. if C1<C2 AND C1<C3 AND C1<C4 AND C1<C5 then
6.2. UPDATE Lc=C1;
6.3. end if;
7. if C2≠0 then
7.1. if C2<C1 AND C2<C3 AND C2<C4 AND C2<C5 then
7.2. UPDATE Lc=C2;
7.3. end if;
8. if C3≠0 then
8.1. if C3<C2 AND C3<C1 AND C3<C4 AND C3<C5 then
8.2. UPDATE Lc=C3;
8.3. end if;
9. if C4≠0 then
9.1. if C4<C2 AND C4<C3 AND C4<C1 AND C4<C5 then
9.2. UPDATE Lc=C4;
9.3. end if;
10. if C5≠0 then
10.1. if C5<C2 AND C5<C3 AND C5<C4 AND C5<C1 then
10.2. UPDATE Lc=C5;
10.3. end if;
11. end if;
12. return Lc;
```

C. Cloud Information Database

Cloud Information database includes the details about the five clouds i.e. Amazon EC2, Cloudsigma, Google, GoGrid, and Windows Azure. It also includes a table of User requirements that saves the input taken by the user. The information database holds the records of the cloud providers

and their details including type of instances, CPU, RAM, storage, standard storage, bandwidth in, bandwidth out, location, operating system, contract period and the pricing details.

I had discussed in the related work section that different cloud providers have their pricing strategies and few of the providers have their own pricing policies which may be totally vary from the rest ones. In Cost aware resource selection system, all the information should be stored in our cloud information database consistently so that the system can display the optimum result without taking help from any other programs. Primary keys have been assigned to every particular provider's table and have explicit details about resources and pricing.

Different combinations have been made in the database for the ease of user to select any kind of combination and find the optimum result.

Following few combinations has been shown using Linear Programming Model (LPM) for better understanding:

$$\begin{aligned}
 I &= \{Ia1, Ia2, \dots, Ian\} \text{ to } \{Iw1, Iw2, \dots, Iwn\} \\
 &\text{(Amazon=a...Windows_azure=w)} \\
 CP &= \{CPa1, CPa2, \dots, CPan\} \text{ to } \{CPw1, CPw2, \dots, CPwn\} \\
 &\text{(Amazon=a...Windows_azure=w)} \\
 R &= \{Ra1, Ra2, \dots, Ran\} \text{ to } \{Rw1, Rw2, \dots, Rwn\} \\
 &\text{(Amazon=a...Windows_azure=w)} \\
 S &= \{Sa1, Sa2, \dots, San\} \text{ to } \{Sw1, Sw2, \dots, Swn\} \\
 &\text{(Amazon=a...Windows_azure=w)} \\
 OS &= \{OS1, OS2\} \\
 Loc &= \{Loc1, Loc2\} \\
 Cn &= \{Cn1, Cn2, \dots, Cnn\}
 \end{aligned}$$

Where $n \geq 1$, I is the set of n instances, CP is the set of n CPU cores, OS is the set of 2 operating systems, R is the set of n RAMs, Loc is the set of 2 locations, S is the set of n Storage, and Cn is the set of n Contract period.

We have introduced few notations to be used in the paper. They are:

I is the instance for every particular cloud, CP is the CPU cores, OS is the type of operating systems, R is the RAM, Loc is the type of location, S is the Storage size, Cn is the Contract period. Let the cloud providers be represented as Amazon_Ec2= a, Cloudsigma=c, Google=g, Go_grid=gg, and Windows_Azure=w.

The database is designed with multiple combinations. The design of the database consists of the fields shown in Table III. The combinations in the database are represented using the linear programming terminologies shown in the following Table II:

The Cost aware resource selection system itself is a Linear Programming (LP) problem. The objective function and the set of constraints are as follows:

$$\begin{aligned}
 &\text{Minimize} \\
 C_{min} &= \sum_{i=1}^n I_n C P_n R_n S_n L_{ocn} O S_n \quad (1)
 \end{aligned}$$

$$\begin{aligned}
 &\text{Subject to} \\
 n &> 0; \quad (2) \\
 L_{ocn} &\in \{1, 2\} \\
 O S_n &\in \{1, 2\} \quad (3)
 \end{aligned}$$

The objective function (1) is the computational cost, which we wish to minimize. With reference to constraint (2), the value of Locn shows whether it can be Loc1 which represents US_east, or Loc2 which represents US_west. Constraint (3) shows that OSn can either be OS1 which represents Linux, or OS2 which represents Windows.

TABLE II. REPRESENTATION OF COMBINATIONS OF DATABASE

Main variables/ Cloud providers	Amazon a	Cloud sigma c	Google g	Go-Grid gg	Windows Azure w
I1	Ia1	Ic1	Ig1	Igg1	Iw1
I2	Ia2	Ic2	Ig2	Igg2	Iw2
:	:	:	:	:	:
In	Ian	Icn	Ign	Iggn	Iwn
CP1	CPa1	CPc1	CPg1	CPgg1	CPw1
CP2	CPa2	CPc2	CPg2	CPgg2	CPw2
:	:	:	:	:	:
CPn	CPan	CPcn	CPgn	CPggn	CPwn
R1	Ra1	Rc1	Rg1	Rgg1	Rw1
R2	Ra2	Rc2	Rg2	Rgg2	Rw2
:	:	:	:	:	:
Rn	Ran	Rcn	Rgn	Rggn	Rwn
S1	Sa1	Sc1	Sg1	Sgg1	Sw1
S2	Sa2	Sc2	Sg2	Sgg2	Sw2
:	:	:	:	:	:
Sn	San	Scn	Sgn	Sggn	Swn
Cn1	Cna1	Cnc1	Cng1	Cngg1	Cnw1
Cn2	Cna2	Cnc2	Cng2	Cngg2	Cnw2
:	:	:	:	:	:
Cnn	Cnan	Cncn	Cngn	Cnggn	Cnwn

TABLE III. FIELDS OF CLOUD INFORMATION DATABASE

Field	Description	Example
Row_id	Primary key to uniquely identify every provider	Row_id=1
Resource (CPU, RAM, Storage)	Resource set of every cloud	CPU= 4, RAM=0.75, Storage=30
Type of Instances	Every cloud has their particular Instances	minimum
No. of machines	Number of cores user requires	50
Location	2 locations are saved i.e. US_east or US_west	US_west
Operating System	2 choices are saved i.e. Linux or windows	Linux
Contract period	It can be hourly, monthly, or yearly	5 hours

D. Results

In Experiment 1, User Selects “Instance based” approach and gives input. In Experiment 2, User selects Quantity based approach and gives Input. The results for both are shown in Table IV. In Experiment 3, User Selects “Instance based” approach and gives input. The results are same as Experiment 1 because this search is based on instance type, and every cloud has their unique instances. No instance of one cloud matches with the instance of other cloud. Hence, in the multi cloud scenario, the results for “Instance based approach” will be same as that of single cloud. In Experiment 4, user selects Quantity based approach and gives Input. Scenarios are taken for Cloudsigma: User gives the input but selects the standard storage of any other cloud other than CloudSigma itself, like standard storage of Amazon EC2, Google, GoGrid, or Windows Azure, to validate that the standard storage can be chosen from any other cloud now.

TABLE IV. EXPERIMENTATION RESULTS

Type of cloud	Approach	Input resources	Optimal cloud	Optimal cost (\$)
Single	Instance based	Type=t2microinstance, No.= 1, Loc=us_east, OS= linux, Contract period= 2hrs	Amazon EC2	0.026
Single	Quantity based	No. of machines=1, RAM (MB)= 0.5, Storage (GB)= 0.5, Contract period= 1, Cloud for storage= Cloudsigma, Standard storage (GB)= 2	Gogrid	0.29
Multi	Instance based	Type=t2microinstance, No.= 1, Loc=us_east, OS= linux, Contract period= 2hrs	Amazon EC2	0.026
Multi	Quantity based	No. of machines=1, RAM (MB)= 0.5, Storage (GB)= 0.5, Contract period= 1, Cloud for storage= Amazon EC2, Standard storage (GB)= 2	Gogrid	0.09

V. CONCLUSION

This paper aims on finding the best resource provider in terms of lowest cost. The system administrators usually spent a lot of time to find the cloud providers with the type of resources they provide and go through the whole pricing policy details. Through cost aware resource selection algorithm, the solution is provided to this particular problem. I reviewed the old approaches where the focus is not on a single factor i.e. cost but also on multiple other factors like efficiency, performance, power, reliability etc. which somehow made the researchers lose the focus on cost. My proposed framework mainly focuses on providing the optimal cost for Single cloud and Multi clouds. I have described the detailed architecture of my cost aware resource selection system and have implemented the prototype. I have deduced some common factors based on our literature review that affected cost in any possible way i.e. Instance type, RAM, CPU cores, computational storage, location, operating system, contract period, number of machines and standard storage. Based on these factors an interface is designed using PHP and java scripting. Input is taken from the user and is matched with that of our cloud information database. Cloud information database is designed in phpMyAdmin to maintain the details of every particular cloud and their pricing policies. The clouds taken for our research are Amazon EC2, Cloudsigma, Google, GoGrid and Microsoft Windows Azure. User requirements are matched with that of the database records and a list is displayed to the user with the costs of all the clouds. Then comparison logic is run to find the optimal and lowest cost among them. This is done under Instance- based approach whereas in Quantity-based approach the results are made through runtime logic. Experimentation and evaluation have proved the validity of the system.

For future work, the parameters that I took for my prototype can be enhanced and increased. Many other locations can be added and more operating systems can be supported in my research. There are many issues that remain open. Something that can search the web interface and updates the cloud information database can be further investigated.

REFERENCES

- [1] Buyya, Rajkumar, et al. "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility." *Future Generation computer systems* 25.6 (2009): 599-616.
- [2] Vaquero, Luis M., et al. "A break in the clouds: towards a cloud definition." *ACM SIGCOMM Computer Communication Review* 39.1 (2008): 50-55
- [3] Mell, Grance. "Peter Mell, Timothy Grance: The NIST Definition of Cloud Computing. Recommendations of the National Institute of Standards and Technology." Gaithersburg, Maryland (2011)
- [4] Motta, Gianmario, Nicola Sfondrini, and Daniele Sacco. "Cloud computing: An architectural and technological overview." 2012 International Joint Conference on Service Sciences. IEEE, 2012.
- [5] Armbrust, Michael, et al. "A view of cloud computing." *Communications of the ACM* 53.4 (2010): 50-58
- [6] Zhang, Miranda, et al. "Investigating decision support techniques for automating cloud service selection." *arXiv preprint arXiv:1210.3077* (2012).
- [7] M. Zhang, R. Ranjan, S. Nepal, M. Menzel and A. Haller, "A declarative recommender system for cloud infrastructure services selection," in Proceedings of the 9th international conference on Economics of Grids, Clouds, Systems, and Services, Berlin, Germany, Springer, 2012.

- [8] Sharma, Upendra, et al. "A cost-aware elasticity provisioning system for the cloud." 2011 31st International Conference on Distributed Computing Systems. IEEE, 2011.
- [9] Liu, Jiangchuan, et al. "A novel framework for QoS-aware resource discovery in mobile ad hoc networks." Communications, 2002. ICC 2002. IEEE International Conference on. Vol. 2. IEEE, 2002.
- [10] Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges." Journal of internet services and applications 1.1 (2010): 7-18.
- [11] Zeng, Wenying, Yuelong Zhao, and Junwei Zeng. "Cloud service and service selection algorithm research." Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation. ACM, 2009.
- [12] Guo, Tian, et al. "Cost-aware cloud bursting for enterprise applications." ACM Transactions on Internet Technology (TOIT) 13.3 (2014): 10.
- [13] Wang, Lijuan, Jun Shen, and Ghassan Beydoun. "Enhanced ant colony algorithm for cost-aware data-intensive service provision." 2013 IEEE Ninth World Congress on Services. IEEE, 2013.
- [14] Grozev, Nikolay, and Rajkumar Buyya. "Inter-Cloud architectures and application brokering: taxonomy and survey." Software: Practice and Experience 44.3 (2014): 369-390.
- [15] Sharma, Upendra, et al. "Kingfisher: Cost-aware elasticity in the cloud." INFOCOM, 2011 Proceedings IEEE. IEEE, 2011.
- [16] Jung, Gueyoung, et al. "Mistral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures." 2010 International Conference on Distributed Computing Systems. IEEE, 2010.
- [17] Alicherry, Mansoor, and T. V. Lakshman. "Network aware resource allocation in distributed clouds." Infocom, 2012 proceedings IEEE. IEEE, 2012.
- [18] Li, Jiayin, et al. "Online optimization for scheduling preemptable tasks on IaaS cloud systems." Journal of Parallel and Distributed Computing 72.5 (2012): 666-677.
- [19] Ferrer, Ana Juan, et al. "OPTIMIS: A holistic approach to cloud service provisioning." Future Generation Computer Systems 28.1 (2012): 66-77.
- [20] Kim, Kyong Hoon, Anton Beloglazov, and Rajkumar Buyya. "Power-aware provisioning of cloud resources for real-time services." Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science. ACM, 2009.
- [21] Le, Kien, et al. "Reducing electricity cost through virtual machine placement in high performance computing clouds." Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2011.
- [22] Lee, Young Choon, Albert Y. Zomaya, and Mazin Yousif. "Reliable workflow execution in distributed systems for cost efficiency." Grid Computing (GRID), 2010 11th IEEE/ACM International Conference on. IEEE, 2010.
- [23] Lakhani, Jignesh, and Padam Kumar. "Resource selection strategy based on propagation delay in Cloud." Communication Systems and Network Technologies (CSNT), 2012 International Conference on. IEEE, 2012.
- [24] Bittencourt, Luiz F., Edmundo RM Madeira, and Nelson LS Da Fonseca. "Scheduling in hybrid clouds." IEEE Communications Magazine 50.9 (2012).
- [25] Dr Ajay jangra, Tushar Saini, "Scheduling Optimization in Cloud Computing", IEEE, Journal paper, Volume 3, Issue 4, April 2013
- [26] Hedwig, Markus, Simon Malkowski, and Dirk Neumann. "Towards Autonomic Cost-Aware Allocation of Cloud Resources." ICIS. 2010.
- [27] Zhu, Mengxia, Qishi Wu, and Yang Zhao. "A cost-effective scheduling algorithm for scientific workflows in clouds." Performance Computing and Communications Conference (IPCCC), 2012 IEEE 31st International. IEEE, 2012.
- [28] Bittencourt, Luiz Fernando, and Edmundo Roberto Mauro Madeira. "HCOC: a cost optimization algorithm for workflow scheduling in hybrid clouds." Journal of Internet Services and Applications 2.3 (2011): 207-227.
- [29] Dixit, Manish Kumar, et al. "Identification of parameters for embodied energy measurement: A literature review." Energy and Buildings 42.8 (2010): 1238-1247.
- [30] Kee, Yang-Suk, et al. "Efficient resource description and high-quality selection for virtual grids." Cluster Computing and the Grid, 2005. CCGrid 2005. IEEE International Symposium on. Vol. 1. IEEE, 2005.
- [31] Jahangirian, Mohsen, et al. "Simulation in manufacturing and business: A review." European Journal of Operational Research 203.1 (2010): 1-13.

ECG Abnormality Detection Algorithm

Soha Ahmed, Ali Hilal-Alnaqbi, Mohamed Al Hemaury and Mahmoud Al Ahmad
United Arab Emirate University
AlAin, UAE

Abstract—The monitoring and early detection of abnormalities in the cardiac cycle morphology have significant impact on the prevention of heart diseases and their associated complications. Electrocardiogram (ECG) is very effective in detecting irregularities of the heart muscle functionality. In this work, we investigate the detection of possible abnormalities in ECG signal and the identification of the corresponding heart disease in real-time using an efficient algorithm. The algorithm relies on cross-correlation theory to detect abnormalities in ECG signal. The algorithm incorporates two cross-correlations steps. The first step detects abnormality in a real-time ECG signal trace while the second step identifies the corresponding disease. The optimization of search-time is the main advantage of this algorithm.

Keywords—Cross-correlation; abnormalities detection; electrocardiogram (ECG); cardiac cycle; eHealth; remote monitoring; algorithm

I. INTRODUCTION

The leading cause for death worldwide is cardiovascular diseases [1]. According to the World Health Organization (WHO), cardiovascular diseases took more than 17.5 million lives in 2012 [2]. Heart attack, hypertension and arrhythmia are among the many heart conditions that fall under the umbrella of cardiovascular diseases [1]. The massive prevalence of heart diseases worldwide urges for novel health solutions. Among the prominent emerging solutions are non-invasive heart rate monitoring technologies and remote health monitoring systems (eHealth systems) [3], [4].

In medicine, Electrocardiogram is a tool used to describe the heart medical condition [5]. Electrocardiogram (ECG) employs multiple electrodes on the chest skin's surface to record the electrical charge resulting from the contraction and expansion of the heart muscle [1], [5], [6]. The analysis of ECG signal can differentiate normal heart beat from irregular one [5]. ECG signal for normal heart beat is uniform and consistent whereas abnormal heart beats have various shapes and forms and each ECG deviation from the normal one is associated with a specific cardiovascular disease [5], [6]. The problem is that standard, 5 to 10 minutes, ECG recording taken at the hospital while the patient is resting may not help the physician identify patient heart problem [1]. Furthermore, the standard procedure is more prone to human error and delayed diagnosis. Thus, a solution to this problem might be realized through the continuous monitoring of patient ECG recording. However, for this solution to be viable, an automated, simple and efficient heart's abnormalities detection algorithms must

be implemented [3]. Those algorithms can be used extensively in eHealth systems and applications and ported to mobiles or sensors processing units [3]. Many ECG classification algorithms were implemented recently [7]. Saritha et al. have used wavelet transform analysis to analyze various cardiac disorders [6]. Furthermore, Rai et al. used artificial neural network classification algorithm to classify ECG signals into two classes: normal, and abnormal [5]. They used two sets of features namely a discrete wavelet transform features and morphological features [5]. They used MIT-BIT arrhythmia database to test their extracted features [5]. The best accuracy they reached was 100% with Multilayer Perceptron (MLP) [5]. Maheshwari et al. developed a novel algorithm to detect the fragmentation of QRS complex in an ECG signal [3]. To verify the validity of their developed algorithm they used PTB database from physioNet [3]. A sample of 31 patients' ECG traces were chosen from the database and annotated by two cardiologists [3]. The algorithm sensitivity was found to be 0.897 and its specificity was found to be 0.899 [3]. Oresk et al. developed a lightweight, real-time cardiovascular disease detection platform for mobiles [1]. The platform they developed employs existing portable ECG monitoring systems and improves it by adding valuable ECG analysis features [1]. They verified the accuracy of their platform using MIT-BIH database. Specifically, they used MLP ANN pattern recognition algorithm to classify ECG signals. They reached a classification accuracy of more than 90% for three abnormal beats. They were also able to identify normal ECG beat with an accuracy of 99% [1]. Shahanaz et al. have proposed a method based on signal processing and correlation technique to find out whether ECG signal is normal or abnormal [8].

In this paper, the proposed algorithm incorporates two cross-correlation steps. The first step is conducted between the real-time ECG signal and the normal stored typical cycle. The role of the first step is to detect abnormality in the real-time ECG signal. In case of abnormality detection, the correlation coefficient value computed in the previous step will be used to reduce the search domain for the second step. Particularly, the second cross-correlation will be carried out between the real-time ECG trace and a small subset of the stored diseases ECG traces. This will reduce the time required for disease detection.

The rest of the paper is organized as follows. Section II describes the proposed algorithm. In Section III, selected cardiac disease and their effect on ECG trace is illustrated. Experiment and simulation results are presented in Section IV. Section V concludes the paper and discusses possible future research directions.

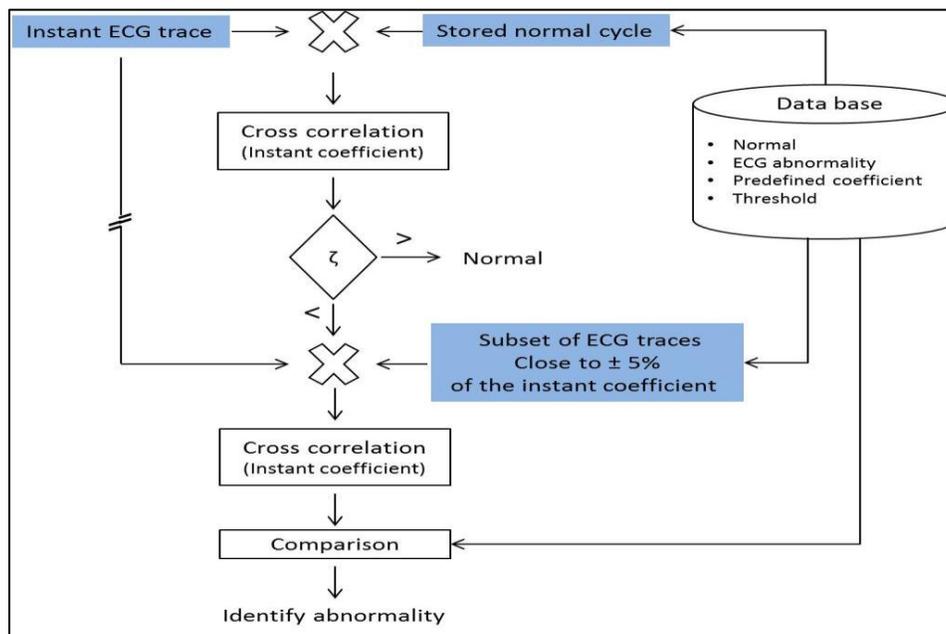


Fig. 1. Proposed algorithm overview.

II. PROPOSED ALGORITHM

The proposed algorithm is illustrated in Fig. 1. The algorithm requires a database that can be modified and updated. This database stores a set of ECG time domain cycles corresponding to different cardiac diseases. In addition, normal cycle ECG signal is also stored to be used as a reference. The correlation coefficients between the normal cycle and the other abnormal cycles are computed and stored as predefined values. The algorithm then computes the instant cross-correlation coefficient between the instant measured ECG cycle and the normal reference cycle stored in the database. If the value of this coefficient is larger than specific threshold (ζ) (in this work the value of ζ is 0.95), then the real-time ECG trace is normal. Otherwise, a mismatch is detected, because the instant coefficient is less than the predefined threshold value ($\zeta=0.95$). This indicates the presence of abnormality in the instant ECG trace.

In case of an abnormality, a second cross-correlation is carried out between the instant ECG trace and a subset of stored diseases ECG traces. The subsets of stored diseases ECG traces have predefined corresponding correlation coefficients close to the instant computed coefficient value by $\pm 5\%$. The resulting correlation coefficient in the second step will identify the cardiac disease.

In other words, the instantaneous ECG trace will correspond to the disease with the largest cross-correlation coefficient value. This procedure reduces the search-time for the corresponding heart abnormality. Particularly, instead of performing the cross-correlation between the instant ECG signal and the whole set of stored abnormal ECG signals; the correlation will only be performed between the instant ECG cycle and a subset of signals corresponding to different cardiac diseases.

III. SELECTED ECG ABNORMALITY CASES

Fig. 2 depicts 7 ECG traces corresponding to well-known heart disease problems. In the following lines, a detailed description of each disease and its effect on the ECG signal will be outlined. The normal ECG signal typical cycle is denoted in Fig. 2(a). Normal ECG signal consists of 6 intervals namely P, Q, R, S, T and U. Each interval represents an electrical incidence during one cardiac cycle. Impulse P is the first short rising of the ECG signal. It implies that the atria are contracting, forcing blood to move into the ventricles. The QRS complex consists of the Q impulse which is downward wave followed by a huge upward wave, namely, R. Then followed by downward wave that is S. The ventricular depolarization and contraction correspond to the QRS complex in the ECG signal. The PR period denotes the transfer time required for the electrical wave to move from the sinus node to the ventricles. Wave T, a small rising of the ECG signal, is signifying ventricular repolarization. Theoretically, the U impulse appears as a result of the repolarization of the interventricular septum. This wave usually has small amplitude, and in most cases, it is entirely absent.

In Fig. 2(b), the ECG signal for a patient with Ischemia is presented. This disease is characterized by inadequate blood and oxygen supply to body organs or more specifically to the heart muscle [9]. The ECG signal for a patient with this disease is distinguished by ST interval disappearance and or downward T wave. The ECG signal for a patient with Injury is depicted in Fig. 2(c). This heart illness is the direct result of having ischemia for a long period of time, which results in the starvation of certain heart tissues for blood and oxygen [9]. The ECG signal for a patient with this disease is distinguished by ST rising. Shown in Fig. 2(d), the ECG signal for a patient with Infraction. Infraction is the death of certain heart tissues as a result of having prolonged Injury [9]. The ECG signal for a patient with this disease is distinguished by the deep Q wave or

the deformed QRS complex [3], [10]. Hypokalemia signifies a low level of potassium in the blood as a result of a certain deficiency [11]. The ECG signal for a patient with this disease (represented in Fig. 2(e)) is distinguished by ST interval disappearance, flat T wave, and appearance of U wave. The ECG signal for a patient with Hypocalcemia is depicted in Fig. 2(f). Hypocalcemia represents a low level of calcium in the blood as a result of deficiency [11]. The ECG signal for a patient with this disease is distinguished by a slender QRS complex, decreased PR segment, and extended ST segment, flattened T wave and noticeable U wave. In Fig. 2(g), the ECG signal for a patient with Hyperkalemia is presented. Hyperkalemia represents a high level of potassium in the blood as a result of deficiency [11]. The ECG signal for a patient with this disease is distinguished by widespread and short amplitude P wave, broad QRS wave, blending of QRS wave with T wave, disappearance of the ST section and elevated slanted T wave. Hypercalcemia ECG signal is presented in Fig. 2(h). Hypercalcemia signifies a high level of calcium in the blood as a result of a deficiency [11]. The ECG signal for a patient with this disease is distinguished by T wave with a wide base and elevated peak.

Fig. 3 will help us demonstrate the proposed algorithm. Fig. 3(a) represents the computed correlation coefficients (ρ) between the normal ECG signal presented in Fig. 2(a) and the other abnormal ECG signals presented in Fig. 2(b) to (h). In Fig. 3(a), the first bar represents the correlation between the normal ECG signal and itself which is typically equals 1. The second bar in Fig. 3(a) represents the cross-correlation coefficient between the normal ECG signal and the ECG signal for a patient with Ischemia and so on. If the signals are identical to each other the coefficient of correlation is unity,

and if they are different from each other the coefficient of correlation is a finite number whose value comes between +1 and -1 inclusive, where 1 represents the total positive correlation, 0 is no correlation and -1 represents total negative correlation [12]. The cross-correlation coefficients presented in Fig. 3 represents a very useful information. They will give us a primary indication of the disease the patient might have. Therefore, these cross-correlation coefficients are saved in the database.

The algorithm usually computes the instant cross-correlation coefficient between the instant measured ECG cycle and the normal reference cycle stored in the database. If the instant coefficient value is greater than or equal 0.95, the real-time ECG trace is normal and the algorithm will continue monitoring the real-time ECG trace. Otherwise, abnormality is detected and the algorithm will fetch ECG trace for diseases with coefficient value $5\% \pm$ the instant coefficient. For instance, if the instant coefficient = 0.52 then ECG trace for infarction ($\rho = 0.56$), hypokalemia ($\rho = 0.54$), and hyperkalemia ($\rho = 0.51$) will be fetched from the database. The second cross-correlation step will be performed between the real-time ECG trace and those three diseases only; this will reduce the search time considerably. The proposed second correlation process between the instant ECG cycle and these three signals at this stage will reveal the identity of the disease. In other words, the instant ECG trace corresponds to the disease whose second cross-correlation coefficient is the largest. In this example, the second cross-correlation coefficient (ρ_2) between instant ECG trace and infarction, hypokalemia, and hyperkalemia equals $\rho_2 = 0.31$, $\rho_2 = -0.08$, and $\rho_2 = 0.97$, respectively. Consequently, the instant ECG trace corresponds to a patient with hyperkalemia.

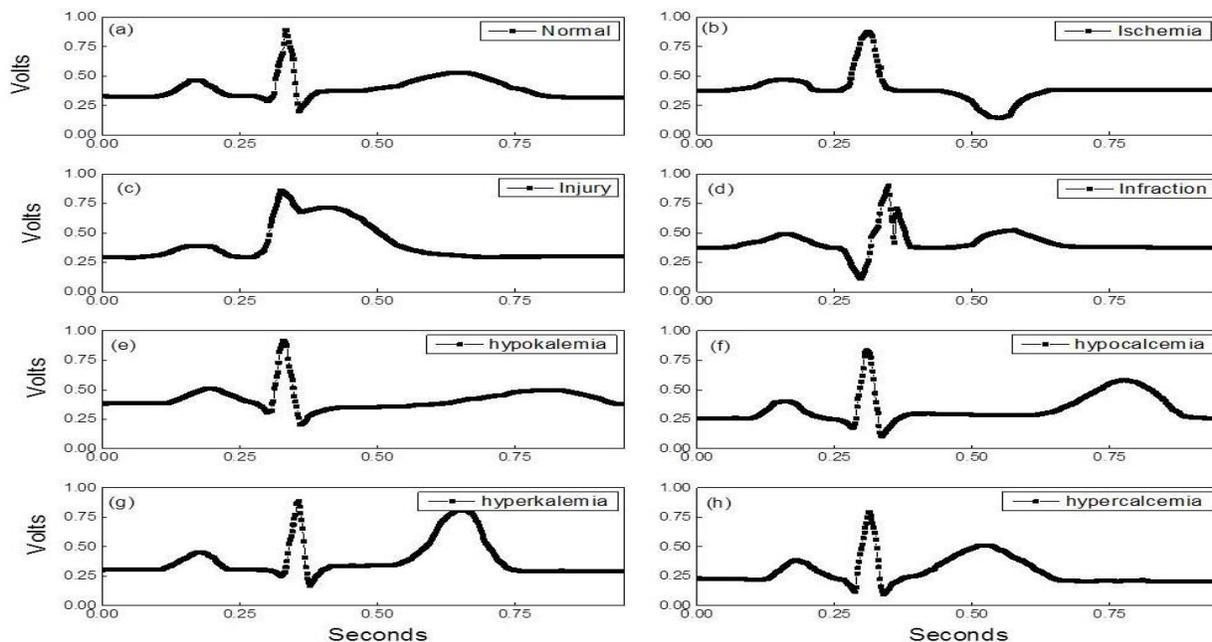


Fig. 2. Normal and abnormal ECG Signals typical cardiac cycle with their corresponding diseases: (a) normal, (b) ischemia, (c) injury, (d) infraction, (e) hypokalemia, (f) hypocalcemia, (g) hyperkalemia and (h) hypercalcemia.

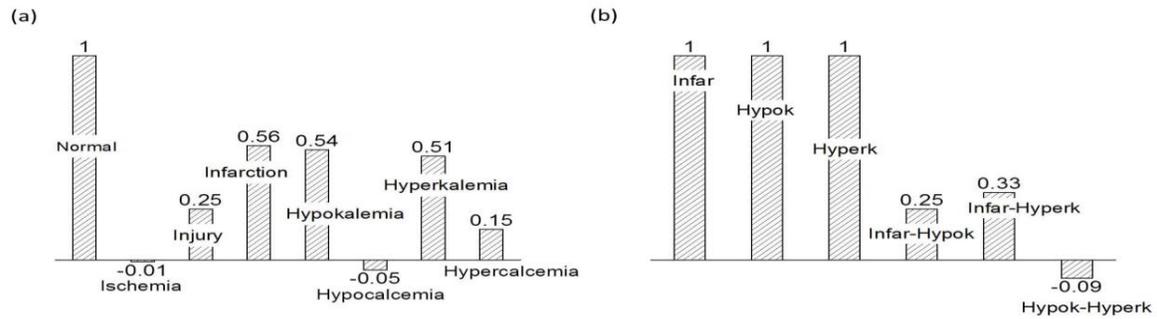


Fig. 3. Correlation coefficients: (a) between normal cycle and cycles of different diseases, (b) between Infarction, Hypokalemia and Hyperkalemia.

IV. EXPERIMENTAL VALIDATION

To validate the proposed approach, the experimental setup shown in Fig. 4 has been utilized. An ECG synthesizer has been constructed. The synthesizer incorporates waveform generator that can be controlled through a software code. The ECG normal cycles were imported through the software code from excel files. Then, the software sends the ECG cycles to the function generator to be repeated continuously. An oscilloscope has been used to display and check the function generator output. The correlation process has been implemented in android tablet application to remotely display the signal and identify any abnormality. The transmission of the instant ECG signal has been carried out via Bluetooth connectivity.

The ECG signals for the seven selected diseases (described in Section IV) have been used for simulation; the detection accuracy was 100%. As expected, in the case of the three mentioned diseases (infarction, hypokalemia, and hyperkalemia) that have close correlation coefficient values, a second correlation process was requested. The second correlation process successfully discriminated each of them with 100% detection accuracy. The incorporation of the second correlation step has been expedited by 40% in each of the seven cases. Assuming the number of stored diseases is 1,000 and 20 out of them were having close predefined correlation values; the speed of the current detection algorithm will then be reduced by 98%.

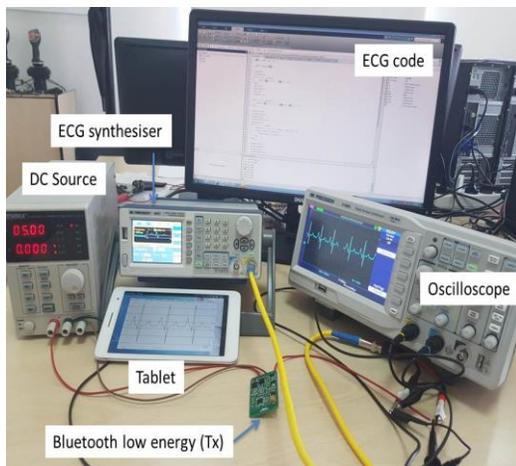


Fig. 4. Real time experimental setup for demonstration.

In summary, a modified correlation based method to detect abnormality in ECG trace has been presented and validated. The proposed algorithm for detecting abnormalities reduces the processing time greatly and the processing power considerably with 100% accuracy.

V. CONCLUSIONS

Naturally, the continuous monitoring of ECG signals will produce large amount of data. Handling this data will require high-performance hardware and software, which will lead to huge energy consumption. Thus, to deal with this issue efficiently, by minimizing resources usage such as the energy and the processing time, an innovative alternative solution should be employed. The deployment of our proposed algorithm will lead to huge reduction in processing time. This processing time reduction stems from the fact that our proposed algorithm compares between numeric values instead of comparing ECG waveforms using image-processing techniques.

This paper addressed the design and implementation of an efficient algorithm. The algorithm detects abnormalities in ECG signal on real-time by utilizing cross-correlation theory. The proposed algorithm for detecting abnormalities reduces the processing time hugely and the processing power considerably.

ACKNOWLEDGMENT

The authors wish to acknowledge the support received from Research Office at the UAE University and ICT Fund UAE.

REFERENCES

- [1] J. J. Oresko, Z. Jin, J. Cheng, S. Huang, Y. Sun, H. Duschl, and A. C. Cheng, "A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 734–740, 2010.
- [2] "WHO | Global status report on noncommunicable diseases 2014," WHO, 2015.
- [3] S. Maheshwari, A. Acharyya, P. E. Puddu, E. B. Mazomenos, G. Leekha, K. Maharatna, and M. Schiariti, "An automated algorithm for online detection of fragmented QRS and identification of its various morphologies," *J. R. Soc. Interface*, vol. 10, no. 89, 2013.
- [4] I. Saadat, N. Al Taradeh, M. Al Ahmad, and N. Bastaki, "Non-invasive piezoelectric detection of heartbeat rate and blood pressure," *Electron. Lett.*, vol. 51, no. 6, pp. 452–454, 2015.
- [5] H. M. Rai, A. Trivedi, and S. Shukla, "ECG signal processing for abnormalities detection using multi-resolution wavelet transform and Artificial Neural Network classifier," *Measurement*, vol. 46, no. 9, pp. 3238–3246, 2013.

- [6] C. Saritha, V. Sukanya, and Y. N. Murthy, "ECG Signal Analysis Using Wavelet Transforms," *Bulg. J. Physic*, vol. 35, pp. 68–77, 2008.
- [7] D. Ge, N. Srinivasan, and S. M. Krishnan, "Cardiac arrhythmia classification using autoregressive modeling," *Biomed. Eng. Online*, vol. 1, p. 5, Nov. 2002.
- [8] S. Ayub BIET, J. Jhansi, and I. J. P Saini, "Abnormality Detection in Indian ECG using Correlation Techniques," *Int. J. Comput. Appl.*, vol. 58, no. 14, pp. 975–8887, 2012.
- [9] R. B. Jennings, "Early phase of myocardial ischemic injury and infarction," *Am. J. Cardiol.*, vol. 24, no. 6, pp. 753–765, Dec. 1969.
- [10] S. Maheshwari, A. Acharyya, P. E. Puddu, and M. Schiariti, "Methodology for automated detection of fragmentation in QRS complex of Standard 12-lead ECG," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2013, pp. 3789–92, 2013.
- [11] D. B. Diercks, G. M. Shumaik, R. A. Harrigan, W. J. Brady, and T. C. Chan, "ELECTROCARDIOGRAPHIC MANIFESTATIONS: ELECTROLYTE ABNORMALITIES."
- [12] K. Callaghan, "The Correlation Coefficient," *Explor. Coll. Algebr. Discov. Databased Appl.*, no. c, pp. 553–557, 1996.

Efficient Resource Consumption by Dynamic Clustering and Optimized Routes in Wireless Sensor Networks

Farzad Kiani

Computer Engineering Dept., Engineering and Natural Sciences Faculty,
Istanbul Sabahattin Zaim University,
Istanbul, Turkey

Abstract—The energy issue is an important parameter in the wireless sensor networks and should be managed in the different applications. We propose a new routing algorithm that it is energy efficient and uses different approaches as dynamic clustering, spanning tree, self-configurable routing and controls energy consuming by data-driven and power management schemas. It has two main phases. The first is consisting of the steady cluster, cluster head election and creation-spanning tree in each cluster and the second phase is data transmission. The proposed protocol is compared with four other protocols in network lifetime, network balance, and average packet delay and packet delivery. Simulation results show the proposed protocol performance in the network lifetime is about 6 per cent higher than Improved-LEACH, 21.5 per cent higher than EESR and 5.8 per cent higher than DHCO. Its improvement in packet delivery parameter is about 3.5 per cent higher than Improved-LEACH, 6.5 per cent higher than EESR and 3 per cent higher than DHCO. In addition, the performance or in packet delay is about 17 per cent higher than EESR and 6 per cent higher than DHCO but Improved-LEACH protocol has a good performance than our protocol about 4 per cent.

Keywords—Energy efficiency; data-driven; spanning tree; sleep/wake up mode; power management

I. INTRODUCTION

The wireless sensor network (WSN) is introduced in mid-20th century [1]. They are one of the distributed systems, subsystems that are in the pervasive systems category. These networks have many applications in medicine, military, smart systems, etc. [2]. So they can be divided to wireless body sensor networks (WBSN), wireless underground sensor networks (WUSN), wireless multimedia sensor networks (WMSN) and such. They are different from the other wireless networks such as Mobile Ad-hoc Network [2]. These networks are combined of large number of mini-size sensor nodes and a few Base Stations (BS) or sink. The nodes have low battery and limited memory. The need of them was felt in many applications and began to spread gradually. Beside the sensor nodes, WSN has one or some of the BS or sink. For example, they can be a computer server. In the network environment, sensor nodes sense phenomenon then collect and process data and send to BS/sink in the end. One of the reasons of development and progression of the WSNs is using the inexpensive and affordable sensor nodes [3]. Therefore, WSNs are used in many applications such as civil, medical, military,

governmental and probability-based applications as volcano [4]. At the beginning, most researchers had focused on bandwidth and Quality of Service (QoS) factors but then energy was considered due to some of the limitations of the networks as battery and memory. Despite the researchers use different techniques to different applications, limitations of WSNs are fix and without change in any application [5]. Hence, one of the most important aims in the WSN is to save energy. The others factors can be different such as QoS and bandwidth so they are the second plan in the network design [1]. There are two kinds of energy consumption between sensor nodes. The first is energy consumption in communications and the second is consumption in computations. The communications consume more energy. Therefore, minimizing communication costs is an important issue. Researchers propose different approaches for this goal such as energy efficiency by routing techniques, data aggregation, duty-cycle techniques, and topology control and medium-access decision.

WSNs are consisting of large number of sensor nodes which the nodes sense data. The data is different due to variant applications and environments [6]. These nodes process and store self-data to memory part after convert sensed physical phenomena to digital signals. Then they send them to the BS/sink via direct or with the aid of other sensor nodes. The scenario is seems perfect but it is not easy according to small size and limited battery of nodes. The nodes are failure prone in real applications. Therefore, management and design of networks is different from other networks. In this mode, lifetime of network will deplete as soon as before correct and complete tasks. It seems that the problem is solved by charge of sensor nodes but charging is difficult or impractical according to concept of WSNs and their applications such as spy network in enemy environment and earthquake or fire prevent networks. Also, energy issue and energy efficiency is very important.

The major goal of some methods is reach to energy efficiency. These methods are based on different approaches such as energy-aware routing protocols or energy efficiency MAC protocols, aggregation of data, sleep/wake up nodes and topology control methods [7]. It should be noted that output parameters implementation of applications are trade off one another. For example, focus on energy parameter can be cause increasing latency or decreasing system reliability [8].

Therefore, always relative balancing should be between parameters. Due to the above reasons, the paper concerns with the energy issue with respect to balancing reasons. The paper is based on increasing lifetime of network by energy-efficiency routing algorithms and energy conservations realize by topology control, data aggregation and sleep/weak up of nodes. The sleep/wake up technique is set of power management schema in energy conservation issues [9]. Fig. 1 shows a view of states node and energy consumption model.

The network architecture has big role in the energy efficiency of these networks. In general, it is based on three models, which are star, mesh and hybrid as shown in Fig. 2. In the star model, a single BS/sink can transmit/receive a message packet to remote sensor nodes. The nodes cannot send the message packets to each other. Home control systems are an example for the model. Their advantages are simplicity, ability to hold down the energy consumption and communications delays between the remote node and the BS/sink. On the other hand, BS/sink must be within all nodes' radio transmission range. In addition, management of the structure has depending to a single sensor node. The last two cases are disadvantages of star structure [11]. In the mesh structure, the sensor nodes can communicate together when they are within radio range of each other. This case realizes multi-hop communication between nodes. Indeed a node can send data to any node (inside self-RF or outside) by intermediate nodes. Scalability and reliability is advantage of the model. If a node is failure then a remote node still can communicate to any other node in its range, which in turn, can forward the message to the suitable place. Energy consumption in multi-hopping system is high generally. Therefore, energy issue is a problem and disadvantage of the model. Moreover, the number of hops and packet delay time increases [11]. Hybrid model is between the star and the mesh structures that provide a robust and self-around communications network. In this model, the sensor nodes with minimum energy are not send message to other nodes and allow to them to saving energy [11].

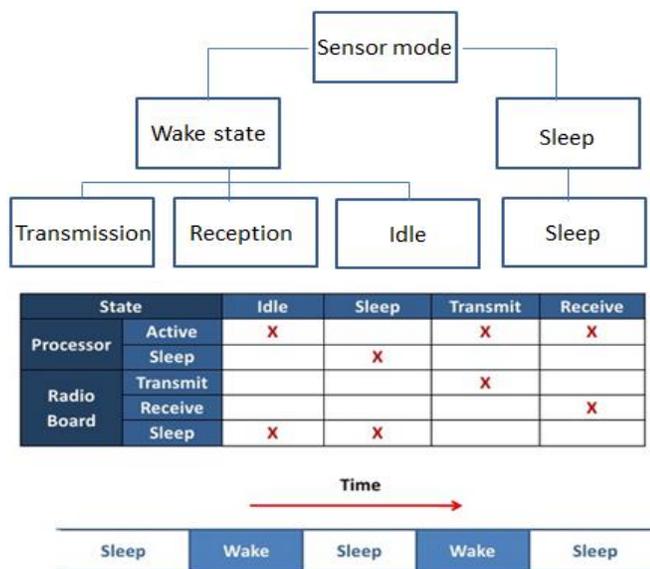


Fig. 1. A view of states node and power consumption [10].

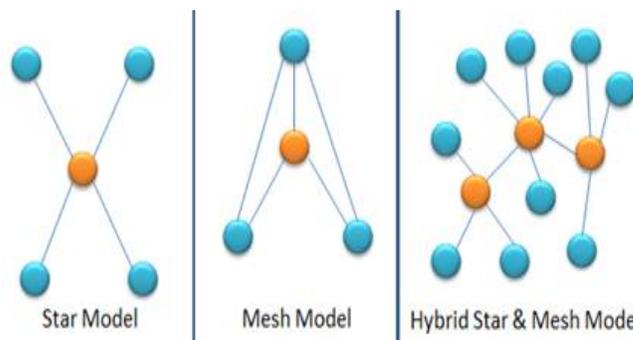


Fig. 2. Structure of a WSN [11].

In Section 2 is discussed the related works about energy efficiency in wireless sensor networks. In Section 3 is introduced a new protocol in order to increasing network lifetime by a dynamic structure that is based on clustering and spanning tree. The results of simulation and evaluation are explained in Section 4. Last section presents the conclusion and future works of the paper.

II. RELATED WORKS

The WSN applications domain are increasing day by day especially in Internet of Things (IoT) based applications [12]. In general, the efficient resource management is essentially issue in these networks. The energy efficiency is converted to popular issue and many of researchers focus on it. Energy efficiency is debatable in all layers of protocol stack. For example, in [13] collision, packet overhead, latency, overhearing and idle listening are discussed and focus on their management to reach energy efficiency. Collisions must be control because they cause unnecessary receive costs at the destination sensor node and cause undue send costs at the source sensor node. The collisions can be managed in the design phase by TDMA and so on protocols and in after the design by avoidance protocols.

In overhearing case is inevitable item in WSNs because the networks have many sensor nodes in an environment and broadcast data transferring is a widespread method. Therefore, possibility of overhearing the nodes within the network is perfectly normal but it is causing increasing energy consumption in the network. Hence, it must be control by some of the approaches as management of nodes density [13], [14]. Latency is gained by delays of transmitted packets in the network. The rate is high in the multi-hop routing protocols.

Idle listening mode is one of the significant reasons increasing energy consumption in the network. When node is active but does not receive any packet or sense any event, it wastes the self-energy extremely. This problem is solvable by different methods as sleep/wake up and MAC protocols. The sleep/wake up scheduling is important case in the protocols. For example, in some of the approaches, the nodes are in sleep mode and upon sense or receive data are changed to active position. In another approaches, the modes changing of each node are depend to time. In methods based on MAC, TDMA, contention based and hybrid schemas manage it. In TDMA-based methods, each node has a time slot and uses it to switching modes. Each of the approaches is discussed in the

following in classification of energy efficiency format. We do not discuss much on the energy-aware data link layer (MAC) protocols. We can say only that MAC protocols are based on content-free and content-based approaches. Content-based approaches have competition for share channel such as MACA and MACAW. In content-free approaches, the channel divides to some sections and each sensor node uses of self-bandwidth without competition. As mentioned in chapter one, some of the content-free approaches are were expressed.

The sensor nodes consume a lot of the energy when they use the control packets. Therefore, the packets number must be managed and the nodes should not use the packets as possible. The packets generally are used in the systems that its goal is reliability. ACK and NACK packets are the samples of the control packets.

According to what was said, energy is one of the most critical resources for WSNs but one problem common to most of them is lack of reliable power for each sensor node in the network. Essentially, data transmission consumes much more energy than data processing. However, the energy consumed by the sensing subsystem varies depending on each node. In some cases, sensing consumes less energy than the one required for data processing while in other cases, it even consumes more than the energy needed for data transmission. In view of the above, several research works has been carried out to solve the energy problem, which results in different schemes and protocols. Most energy conservation techniques in the networking and sensing subsystem is proposing energy efficient protocols to minimize energy consumption during network activities and power management schemes for switching off idle node components are necessary for maximum energy conservation in wireless sensor networks.

We focus on energy efficiency issue in this paper and one of the main our goals is energy saving and prolonging network lifetime. For example, chapter four will propose a new routing algorithm with rely on data-driven and sleep/weak up of nodes techniques. The chapter five will explain a new routing protocol base on machine learning technique that use of data-driven technique of energy efficiency schemas. The chapter six will introduce a novel routing algorithm based on topology control technique. All the techniques are discussed in this chapter completely.

We can classify energy efficient schemes and protocols in WSNs. They are into three classifications so duty-cycling, data-driven and mobility-based methods. Duty cycle schema focuses on subsystem networks and radio transmission switching. Main work of duty-cycle base approaches is maintenance radio transceiver in low power state by sleep mode and it is realizable whenever a sensor node does not communicate with other nodes. If a node is idle and does not senses/sends/receives, the radio mode of the node will wake up to energy consumption management.

Process unit of sensor node do exchanging sleep to wake up mode or vice versa in special and defined periods. This task is done a sleep/wake up scheduling algorithm within any protocol based on duty cycle schema. It is typically a distributed algorithm based on which sensor nodes decide when to transition from active to sleep and back. It allows neighboring

nodes to be active at the same time, thus making packet exchange feasible even when nodes operate with a low duty cycle (i.e., they sleep for most of the time). Duty-cycling schemes are typically oblivious to data that are sampled by sensor nodes. On one hand, data-driven approaches are the other method of energy efficiency that can be used to improve the energy saving even more. Data sensing impacts on sensor nodes' energy consumption are in two topics. Sampled data generally has strong spatial and/or temporal correlation [15] so there is no need to communicate the redundant information to the sink. In fact, they are unnecessary samples. The reduction is not enough when the power of self-sensor is low. This issue arises whenever the consumption of the sensing subsystem is not insignificant. Data driven techniques presented in the following are designed to reduce the amount of sampled data by keeping the sensing accuracy within an acceptable level for the application. In case some of the sensor nodes are mobile, mobility can finally be used as a tool for reducing energy consumption (beyond duty cycling and data-driven techniques). In a static sensor network packets coming from sensor nodes follow a multi-hop path towards the sink(s). Thus, a few paths can be more loaded than others can, and nodes closer to the sink have to relay more packets so that they are more subject to premature energy depletion (funneling effect) [16].

If some of the nodes (including, possibly, the sink) are mobile, the traffic flow can be altered if mobile devices are responsible for data collection directly from static nodes. Ordinary nodes wait for the passage of the mobile device and route messages towards it, so that the communications take place in proximity (directly or at most with a limited multi-hop traversal). Consequently, ordinary nodes can save energy because path length, contention and forwarding overheads are reduced as well. In addition, the mobile device can visit the network in order to spread more uniformly the energy consumption due to communications. When the cost of mobilizing sensor nodes is prohibitive, the usual approach is to "attach" sensor nodes to entities that will be roaming in the sensing field anyway, such as buses or animals. The classification is shown in Fig. 3 with detailing of subgroups [17].

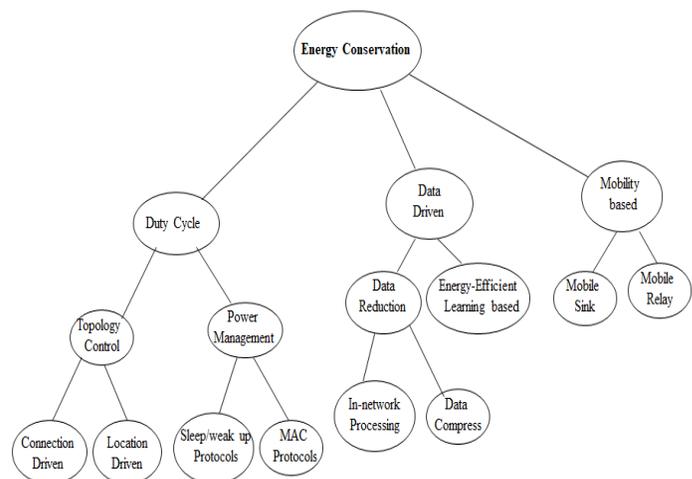


Fig. 3. Classification of energy efficiency schemas in WSNs [7].

III. PROPOSED SYSTEMS

We propose a new routing algorithm to optimize energy consumption in the WSNs. The protocol is based on a hierarchical approach and it uses a tree structure for deploying network and routing on it. We show that our algorithm has the improvement rather to some of the current methods such as LEACH [15], Improved-LEACH [16], EESR [17] and DHCO [18] in similar fields. Our protocol (EESTDC) has two main phases. The first phase is consisting of steady cluster, Cluster Head (CH) election and Spanning Tree (ST) creation in each cluster and the second phase is data routing. Our protocol is based on a dynamical model in CH selections, changing topologies structures in any round of running network. This can reduce overhead and improves resources consumption in the whole system.

Structure of communications between sensor nodes in the clusters is based on spanning tree. The tree is structured in the way that the node with the smallest identifier is chosen as the root. All other nodes are connecting to this selected root via the shortest-path route. CH election is based on specific method that uses the residual energy parameter. In the method, overhead over a CH node will be reduced because CH nodes are changed in per round. Routing phase in our algorithm is based on spanning tree algorithm. In fact, it uses a new approach for data transferring from BS/sink to other sensor nodes or vice versa. Tree structure is applied to every cluster after determination CH nodes. This architecture helped data aggregation in each layer. Therefore, data aggregation task did not impose on the CH nodes. In addition, it used the TDMA [19] technique in data transferring phase. For example, if a cluster is consisting of 10 sensor nodes (9 nodes and one CH node) then CH will cut bandwidth between nine nodes by the TDMA.

CH election method is shown in the following formula. The first part of this formula is similar to CH election in the LEACH protocol. Heavy duty imposed on a CH node is an important weakness in the LEACH and it does not change CH node in the whole network lifetime. These problems are solved by added parts to formula in this case that is shown in (1).

$$\begin{aligned} T(I) &\leftarrow P(I) \div (1 - P(I) * (r \% (P(I)))) \\ CH(I) &\leftarrow T(I) \div E(I) \\ CH(R) &\leftarrow \min(CH(I)) \end{aligned} \quad (1)$$

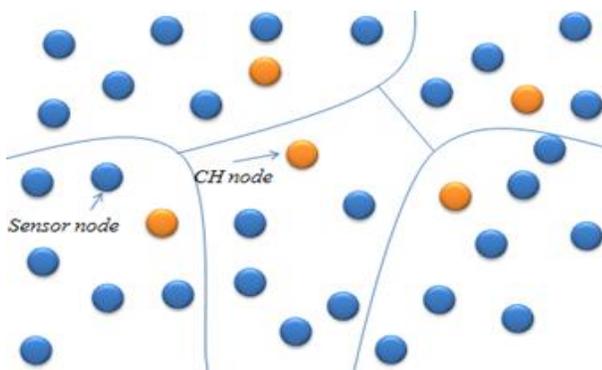


Fig. 4. An example of network after selection CH Nodes.

In this case, P is a random number that identifies percent of CH node possibility. R show current round and T (I) is a threshold that its value is between zero and one. This value calculates for all sensor nodes in every cluster. I is number of each sensor. E (I) is residual energy of i^{th} node. In finally, a node is selected as CH node that has minimum value among CH (I). This model will cause to reduction energy consumption on CH nodes, energy balance of the whole network and prolonging lifetime of network. Fig. 4 shows an example of clusters formation and selected CH node to every cluster.

After election CH node process, selected CH node broadcast an announcement message to neighbors` nodes that it had selected as CH node. Selected node is a CH node for one round only. After receiving message by the each sensor node, it decides be a member of a cluster so it will transmit its data to CH node of the cluster. If a sensor node has same distance from two or more CH node then it selects a cluster by random. Tree structure is applied to every cluster after determination CH nodes. This architecture helps data aggregation in each layer. Therefore, data aggregation task does not impose on the CH nodes only. This will cause to reduction of system overhead and increasing network lifetime. The architecture has variable topology due to changing in CH nodes in each round. The nodes must be configured in the range new CH nodes for reach to acceptable level in the re-organization of tree in the each round. Hence, not all structure of the tree changes in different rounds of running network. The creation of tree is an iterative procedure. Every sensor node is located within a cluster and has a relation with its CH node. This case is done by tree spanning model. Indeed, this model makes all communications between nodes and nodes-CH nodes. Each node selects its children and this case is iterative to end. In the end, all sensor nodes are within a tree structure so this structure is based on spanning tree algorithm. After the determination of communication channel of each node, every one of them is active for a few seconds and then is changed to sleep mode. This is a method for energy efficiency in the WSNs so it can avoid unnecessary consumption of energy in each node and this will cause to prolonging lifetime of the network. A sensor node has three main parts as sensing unit (S.U), processing unit (P.U) and transmission unit (T.U) [2]. Generally, T.U and S.U can sleep but P.U is always on in the all conditions. When S.U goes to sleep mode, it buffers sensed data so it can send them after changing mode to wake up. The waking of the S.U is task of the P.U. For example, sensor is activated once every ten seconds by the P.U. In the sleep mode, memory part of P.U may be in sleep mode for more energy saving but it must wakes up after activation of S.U. On the other hand, T.U cannot send or receive any data packet of its neighbors or BS/sink in sleep mode. After activation of node`s RF, it can send the buffered data to its neighbors [8]. The crucial point is that while T.U is sleep mode, node never cannot receive or send data and this case is problematic especially in the target tracking applications. One of the good transmitter modules for short-range schemas is TR1000 module. It has a short range but energy consumption in receiving part is quarter of send part. Indeed, we can hold off the portion of the reception to receive possible message packets of neighbors. Therefore, they can receive wake up or target tracking messages of neighbors. This case is a semi-sleep model and it can active reception part

of T.U. Our protocol uses the module because it is suitable in target tracking applications such as possible prosecution of enemy tanks. Communications between nodes within each cluster are hierarchical-based and transmissions are from down (child node) to up (parent node). The data packets are aggregate by each the receiver nodes. Hence, CH node will receive low volume data packet of children and this will cause to prolonging the survival rate of the network. In fact, our protocol increases network lifetime and is energy efficient in three visions. The CH nodes will be available more time and they will have long-lifetime in the network. This is the result of two reasons. First, reduction tasks over CH node by data aggregation technique in per nodes. The second is repeated changes in clusters of heads task. Second vision is sleep/wake up approach that uses a specific module for our application types. Third vision is sleep mode in some nodes that they are unrelated to the target. Fig. 5 shows an example of executable schema of EESTDC after finish second phase in a random current time. Fig. 6 represents a pseudo code for set-up and data transferring phases of protocol.

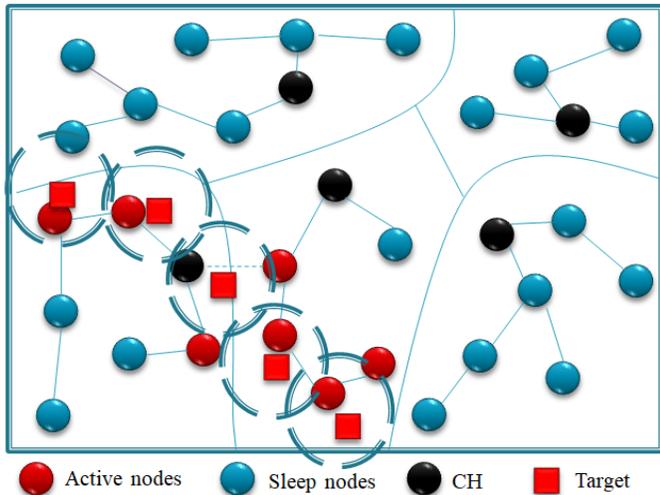


Fig. 5. Example schema from operation in target tracking application after termination phases (CH Node is Active Mode).

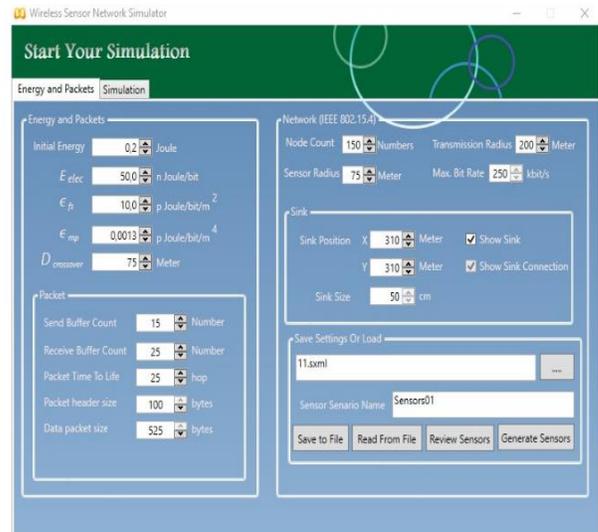
```

// Variables
T: threshold value, r: round of network, i: node numbers. E: energy of nodes,
P: random number [0, 1]
// clustering and Election of CH nodes
Network is divided to several clusters
While (r is true)
Select CH node to each cluster
T (i) = P (i) / (1 - P (i)) * (r mod (1 / P (i)))
CH (i) = T (i) / E (i), CH (r) = Min (CH (i))
//Broadcasting by selected CH node
CH node broadcast his success in the CH election message
// to recruits members' nodes to clusters
Each message receiver node chooses to be a member of cluster. This decision
is apply by some parameters such as distance it to CH node. Node chooses the
nearest CH node.
If distance of two or more CH node is same
Then node select one of them randomly
// Building spanning tree
S=number nodes of a cluster
M=1, N=1, CH=root
While (cluster number>0)
List-level (1) =neighbors CH, Queue (M) =list-level (1), M=M+1, S=S-1
While (S>1)
List-level (N+1) =neighbors (list-level (N)), Queue (M) =list-level (N)),
M=M+1, S=S-1
End While
End While
Spanning tree is to finding shortest route from CH node to all nodes of it
cluster
While (CH receive all INFO)
Nodes aggregate self-data and send INFO to self-neighbors nodes
End While
    
```

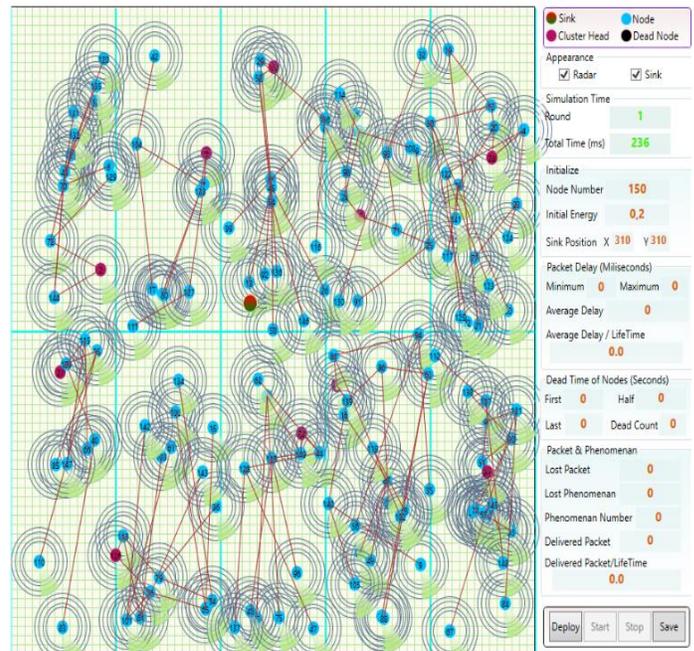
Fig. 6. Overview of pseudo code for setup and data transmission phases.

IV. RESULTS AND EVALUATION

We made a WSN simulation tool in C# program that uses it for all simulations in this paper, as shown in Fig. 7. The tool allows us to have results documentation or simulation charts in network lifetime, packet delivery and packet delay parameters at the moment or end. In this case study, we simulate all protocols with the same outputs parameters too such as network lifetime, packet delivery, packet delay and network balance. In addition, we use the original simulation charts of Improved-LEACH, DHCO, and EESR to demonstrate the correctness of the proposed protocol. We apply their input parameters for EESTDC to comparison and then will use our tools for further simulations.



a) Simulator Setting



b) Snapshot of simulator program

Fig. 7. A snapshot of basic network simulator. a) Simulator setting, b) Work of the program.

TABLE I. VALUES OF INPUT PARAMETERS FOR EESTDC, IMPROVED-LEACH, EESR AND DHCO PROTOCOLS

Initial (max) energy	0.2 J/bit	Receive buffer size	1500 bytes
Radio/ Sensor energy consumption	50 n J/bit	Send buffer size	1500 bytes
E_{elec}	30 n J/bit	Deployment area size	(600 x 600) m
ϵ_{fs}	10 p J/bit/n ²	Send/receive buffer counts	10
ϵ_{mp}	0.0013p J/bit/n ⁴	$D_{crossover}$	75m
Data packet size	150 bytes	Sink position	(310 x 310) m
Sensing Radius	4.5m	Transmission Radius	9m

The complexity of our work is $O(cnd)$ that c is packets amount, n is number of network nodes and the d indicates the number of neighbors of each node. We assume the input parameters values have listed in Table I. We ran each of protocols in seven cases with different node numbers. As it seems, EESTDC has a good performance in network lifetime and can increase this factor by methods such as sleep/wake up, data aggregation, applying the new approach in CH node election and spanning tree methods. Fig. 8 illustrates the results of all four protocols in terms of network lifetime. EESTDC has a good performance than other protocols. The EESTDC improvement is about 6 percent higher than Improved-LEACH, 21.5 percent higher than EESR and 5.8 percent higher than DHCO.

The second case of comparison is packet delivery. As mentioned, the number of successfully transmitted packets from a node to sink and their reception by the sink is the concept of packet delivery. In addition, packet loss is gained from subtraction of all sensed data packets and number of delivered packets. The proposed protocol has low optimization in packet delivery than Improved-LEACH, DHCO, and EESR unlike its improvement in the network lifetime. The improvement is about 3.5 percent higher than Improved-LEACH, 7.5 percent higher than EESR and 3 percent higher than DHCO. Fig. 9 shows the packet delivery rate for all protocols.

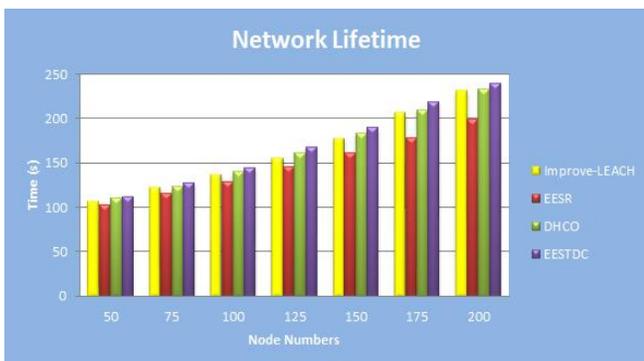


Fig. 8. Network lifetime simulations with different node numbers in improved-LEACH, DHCO, EESR and EESTDC protocols.

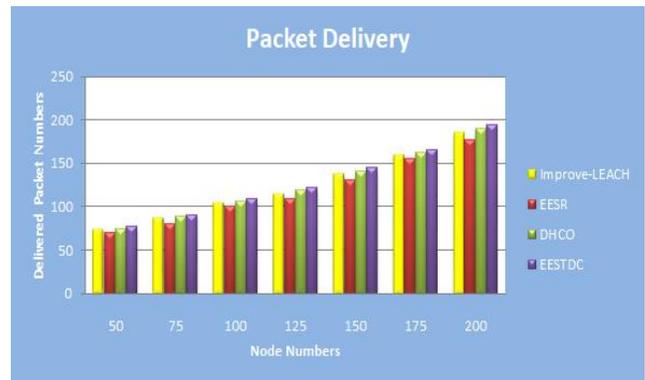


Fig. 9. Network lifetime simulations with different node numbers in improved-LEACH, DHCO, EESR and EESTDC protocols.

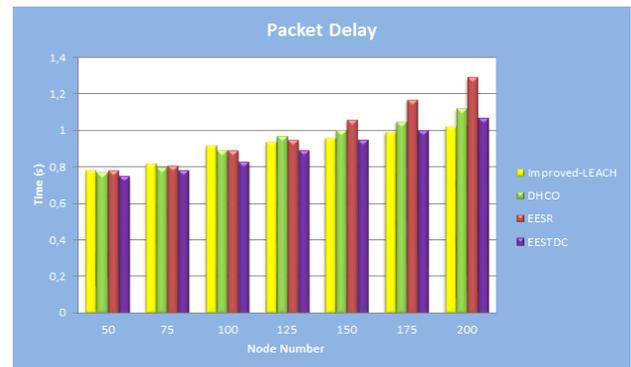


Fig. 10. Packet delay simulations in different node numbers cases for improved-LEACH, DHCO, EESR and EESTDC.

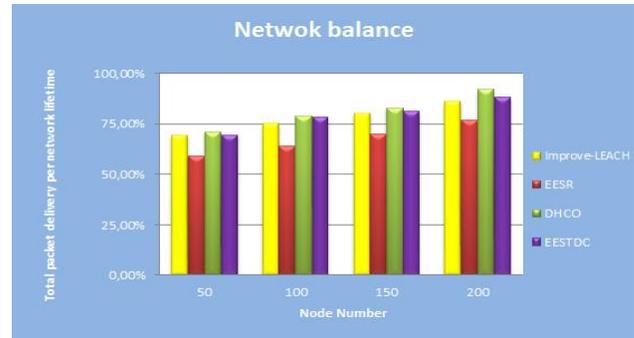


Fig. 11. A specific view from performance of improve-LEACH, DHCO, EESR and EESTDC in network balance.

The third case of comparison is packet delay. As mentioned, packet delay is a period time that a transmitted packet will reach to BS/sink in this time. In fact, a transmitted packet will consume sometimes to reach to BS/sink. This time is a delay for every data packet. Fig. 10 illustrates the simulation results for all four protocols. As it seems, EESTDC is second suitable approach after Improved-LEACH protocol. Increasing the delay rate in EESTDC has balance.

The last simulation parameter is the network balance. As mentioned, network balance is another measure to the simulation of four protocols. If the main goal of network design is increasing reliability, network lifetime will reduce automatically. With this description, if a protocol can make a balance between output parameters then it would have a good

REFERENCES

performance in general views. It can gain from different methods such as percent of packets delay to network lifetime, percent of packet delivery to network lifetime, etc. Fig. 11 shows a case of network balancing comparison that it is calculated based on the relation between delivered packets numbers and network lifetime. Our expectation from EESTDC is not very good performance because reach to ideal network balance is actually impossible; although, the proposed protocol has an acceptable balance level.

V. CONCLUSION AND FUTURE WORKS

As mentioned, energy is an important parameter in the WSNs, hence, the proposed protocol must manage it in the various conditions. Routing algorithms can reduce energy consumption in sensor nodes by finding optimum routes from source nodes to sink and between all nodes. Significant approaches for realizing energy efficiency are hierarchical based routing algorithms that the most applications can use them to reach the goal, however, they have some of the problems such as does not have an optimal network lifetime continually or have overhead over CH nodes. Each of the methods had some of the advantages and disadvantages, as they were usable in special applications only. Simulation results show the proposed protocol performance in the network lifetime is about 6 per cent higher than Improved-LEACH, 21.5 per cent higher than EESR and 5.8 per cent higher than DHCO. Its improvement in packet delivery parameter is about 3.5 per cent higher than Improved-LEACH, 6.5 per cent higher than EESR and 3 per cent higher than DHCO. In addition, the performance or it in packet delay is about 17 per cent higher than EESR and 6 per cent higher than DHCO but Improved-LEACH protocol has a good performance than our protocol about 4 per cent.

The proposed protocol can realize energy saving and be prolonging lifetime factors. In addition, EESTDC has suitable performance in the packet delivery parameter. However, levels of network balance and packet delay of EESTDC are medium. The other of the main shortcomings of our protocol is a lack of focus on relationships between the nodes of different clusters. Meanwhile, it seems that CH node election and restructuring tree in per round of running network has extra overhead over the system. In addition, it seems that an intelligent approach can solve the disadvantages automatically and it tries to keep a balance in the whole of the network.

As future works, we offer some of the suggestions in the fields. So, one of them is focus on the communication between CH nodes in intra-network and optimize their relations. The second is using the learning based routing techniques and combining with the proposed method. Lastly, use the fuzzy logic methods to create optimal communication paths and activation of nodes can be proposed.

ACKNOWLEDGMENT

This project has been achieved in Wireless Sensor Networks and IoT Laboratory in Istanbul Sabahattin Zaim University (IZU). The authors declare no conflict of interest.

- [1] A. Kumar, T. Sato, "Localization in Wireless Sensor Networks: A Survey on Algorithms, Measurement Techniques, Applications and Challenges", *Journal Sensor and Actuator Networks*, 6, pp. 1-23, 2017.
- [2] F. Kiani, E. Amiri, M. Zamani, T. Khodadadi, and A. Abdul Manaf, "Efficient intelligent energy routing protocol in wireless sensor networks", *International Journal of Distributed Sensor Networks*, 2015, pp.1-13, 2015.
- [3] M. Ouadou, O. Zytoune, Y. Hillali, A. Menhaj, D. Aboutajdine, "Energy Efficient Hardware and Improved Cluster-Tree Topology for Lifetime Prolongation in ZigBee Sensor Networks, *Journal Sensor and Actuator Networks*, 6, pp. 22-37, 2017.
- [4] V. Ramasamy, "Mobile Wireless Sensor Networks: An Overview", *Open Access Peer-Reviewed Chapter, Books/Wireless Sensor Networks*, pp. 3-19, 2017.
- [5] I. F. Akyildiz, Y. Sankarasubramaniam, E. Cayirci, "A survey on sensor networks", *IEEE Communications Magazine*, 40(8), pp.102-111, 2002.
- [6] F. Kiyani, H. Tahmasebirad, H. Chalangari, S Yari, "DCSE: A dynamic clustering for saving energy in wireless sensor network", *IEEE International Conference on Communication Software and Networks*, pp. 13-17, 2010.
- [7] G. Anastasi, M. Coti, M. Francesco, A. Passarella, "Energy conservation in wireless sensor networks: A survey", *Elsevier, Ad Hoc Network*, 12, pp. 537-569, 2009.
- [8] F. Kiani, A. Aghaeirad, M. K. Sis, A. Kut, , A. Alpkocak, "EEAR: An Energy Effective-Accuracy Routing Algorithm for Wireless Sensor Networks", *Life Science Journal*, 10(2), pp. 39-46, 2013.
- [9] S. LAI, "Duty-cycled wireless sensor networks: Wakeup scheduling, routing, and broadcasting", PhD Thesis, Virginia Polytechnic Institute and State University, 2010.
- [10] M. Arifuzzaman, M. Matsumoto, "An Efficient Medium Access Control Protocol with Parallel Transmission for Wireless Sensor Networks", *Journal Sensor and Actuator Networks*, 1, pp.111-122, 2012.
- [11] H. Khanmirza, "Mitigating energy hole problem with power control in heterogeneous sensor networks", *IEEE Conference, 2017 Iranian Conference on Electrical Engineering (ICEE)*, pp. 736-741, 2017.
- [12] F. Kiani, "A Survey on Management Frameworks and Open Challenges in IoT", *Wireless Communications and Mobile Computing*, 2018, pp.1-33, 2018.
- [13] M. C. Vuran, O. B. Akan, I. F. Akyildiz, "Spatio-temporal correlation: Theory and applications for wireless sensor networks", *Computer Networks Journal*, 45(3), pp. 245-259, 2004.
- [14] F. Kiani, Animal behavior management by energy-efficient wireless sensor networks, *Elsevier Journal, Computer and Electronic in Agriculture*, 151, pp. 478-484, 2018.
- [15] W. Heinzelman, A. Chandrakasan, H. Balakrishnan, "Energy-efficient communication protocol for wireless micro sensor networks", *International Conference on System Sciences*, pp.1-10, 2000.
- [16] J. Y. Lee, K. D. Jung, S. J. Moon, "Improvement on LEACH protocol of a wide-area wireless sensor network", *Multimedia Tools and Applications*, 76, pp. 19843-19860, 2017.
- [17] S. Hussain, O. Islam, "An energy efficient spanning tree based multi-hop routing in wireless sensor networks", *International Conference on Wireless Communications and Networking (ACM)*, pp. 4386-4391, 2007.
- [18] Y. Chang, H. Tang, Y. Cheng, Q. Zhao, B. Li, X. Yuan, "Dynamic Hierarchical Energy-Efficient Method Based on Combinatorial Optimization for Wireless Sensor Networks", *MPDI Sensors*, 17, pp. 1665-1680, 2017.
- [19] T. Kaur, D. Kumar, "TDMA-Based MAC Protocols for Wireless Sensor Networks: A Survey and Comparative Analysis", *5th International Conference on Wireless Networks and Embedded Systems (WECON)*, pp. 1-6, 2016.

A Method of Automatic Domain Extraction of Text to Facilitate Retrieval of Arabic Documents

Mohammad Khaled A. Al-Maghasbeh, Mohd Pouzi bin Hamzah
School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu
Kuala Terengganu, Malaysia

Abstract—Arabic content on the internet has increased over the web because of the growth of the number of Arabic persons who use the internet in the world. Accordingly, this study introduces an automatic approach of domain extraction of information retrieval from these contents based on text classification. Text classification process makes the searching domain specific to facilitate the searching process. This paper discusses how to enhance the capacity of information retrieval in Arabic documents by classifying the unlabelled Arabic text automatically by using text classification algorithms. The classification of documents and texts is an important field in computer science and information retrieval. It aims at enhancing the retrieval process by identifying the searching-domain of retrieval systems.

Keywords—Arabic information retrieval; text classification; Arabic text mining; Arabic language processing; text clustering; text classification; text categorization and classification algorithms

I. INTRODUCTION

The Arabic language is one of the most common languages spread over the world which represents as one of the natural languages used in information retrieval field. Arabic language can be classified into three categories of dialects according to use in our life: traditional language, the formal language “Classical” and modern standard language. The first type is the language spoken or colloquial Arabic among people in their life, the second type is Holy Quran language written, and the third is a common language in the Arab world that commonly being used in literature, poetry, stories and literary writings [5].

Despite increasing the Arabic documents over the web. There are many problems still makes the Arabic information retrieval so challenge. The main problem in Arabic information retrieval is how to improve the retrieval accuracy. Hence, in this study, the new approach is proposed to merge the two applications of the NLP, namely, information retrieval and text classification.

This rest study is organized as follows. Section 2 briefly describes the related works in the area of Arabic texts classification. Section 3 describes Arabic text classification. Section 4 provides proposed approach. Section 5 shows the discussion with an example. In Section 6, it summarizes the work and future work.

II. RELATED WORKS

There are many applications of information retrieval. One of these applications is document classification that involves classifying document or text into several categories depends on some factors. Fraud et al., in their study used one of the information retrieval, namely, latent semantic analysis model with five similarity measurements to enhance the Arabic documents clustering. LSA in this study has been applied by using Singular value decomposition (SVD) to create an abstract representation of each document [9].

Mohammad Naji applied six common text classification methods such as Naïve Bayesian method (NB), support vector machines (SVM), Rocchio algorithm, k-Nearest Neighbor (KNN), neural network (NNet), Linear Least Squares Fit (LLSF) on two of set of Arabic document for training and test towards build a system to obtain a similarity degree between the training and test sets vectors based on the inner product feature. After that the proposed system computes the cosine between two vectors to find the best or appropriate class for each document has been tested. The recall and precision were the best with high similarity degree in Naïve Bayesian method [2].

In Thabtah study, the text classification has achieved based on Naïve Bayesian model with uses the mathematical statistics method that measures the correlation between two variables to check either correlated or independent. This mathematical method is known as the Chi-square method. The dataset in this study were 1562 Arabic documents from Sudia Press Agency (SPA) that is classified into six categories (Social, Cultural, Sports, Political, General, and Economic) [15].

The El Kourdi, study concerned for automatic classification of the Arabic web documents to help the search engine to deal with the continuous growth of the document via the internet. The Naïve Bayesian (NB) applied in this study to classify 300 of the Arabic web document that taken from Al-Jazeera website “the channel of Arabic News in Qatar Television” into five categories “Science, Health, Culture and Art, Business, and Sport”. The results showed accuracy in classification reaches 92.8% while the manual methods reached 62.8 [7].

Ababneh, in his study has been applied a Support Vector Machine algorithm (SV) on 5121 Arabic documents from the Saudi Newspaper (SNP) using K- Nearest Neighbor (KNN) technique to classify it’s into seven classification categories includes “Economics, Information Technology, General,

Politics, Cultural, Sports, and Social". In this work, the Arabic documents classified depend on the similarity degree, whereas it has been applied a several of experiments in (KNN) and (SVM) by using three different coefficients (Jaccard, Cosine, and Dice) to do a comparison between of them. This work took the measures F1, Recall, and Precision to compute the efficiency of the two algorithms. The method has been proved that the Cosine was better than Jaccard, and Dice coefficients [1].

A study by [8] carried out some experimental for classification the Arabic texts from newswire by using mathematical or statistical classification methods. The experimental dataset was in sport, economics, politics, and culture, whereas 80% of these data used for training, and the 20% for testing

III. TEXT CLASSIFICATION

There are more than one ways that used in text mining to classify the documents into groups to represent the knowledge from them. These methods use set of factors to classify the number of documents into classes based on similarity, subject, and other characteristics [2]-[14]. Hence, there many traditional algorithms that use to classify the texts such as K-Nearest Neighbor (KNN), Naïve Bayes model, Decision tree, Support vector machine (SVM), and artificial neural networks (ANN) [10].

Text collection (TC) is a process to classify the document into groups based on the similarity, and it applies in some applications such as text filtering, web page categories, document organization, and other [11], [13]. The text classification has been become importance due to increasing the amount of data such as stories, news on the internet to facilitate the task of information retrieval when needed [15].

IV. PROPOSED APPROACH

We propose a new approach for information retrieval from Arabic documents depends on texts, or documents classification task as shown in Fig. 1. In this approach, we use the keywords extraction documents to classify them into several categories, and then in user query will extract the query terms. This model makes the system able to compute the similarity between the terms of the query with related documents through determining the domain in both, documents and user query.

The mechanism of the proposed approach has explained at the pseudo code that mentioned as follows:

Input: Documents collections, and General query;
Output: Retrieve the document that relevant the query;
For each query;
Begin:
Extract the query terms;
Compute the similarity between the query terms;
Determine the query domain;
Go to indexed documents, and then:
Extract all candidate keywords from each document;
Compute the similarity between document keywords;
Classify the document into categories;
Indexing each document category;
While similarity (Query domain, Doc domain) do:

Search in all the document in the same category;
Match (Query Term, Document keywords);
Retrieve all of related documents;
End

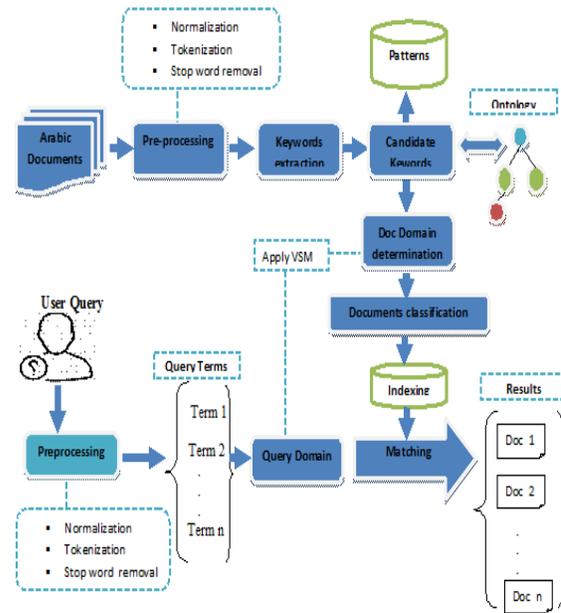


Fig. 1. Architecture of proposed approach.

The proposed model includes two main phases; each phase includes subtasks as the following:

1) In documents processing

- The pre-processing phase that includes the normalization, tokenization, and stops words removal.
- Extract the keywords in each document using the ontology, and the patterns that saved from other documents.
- Extract general topic or domain through computing the vectors space between the document keywords using VSM.
- Classify the documents based on Cosine similarity.

2) In user query processing

- Apply the preprocessing techniques like document preprocessing.
- Extract the query terms.
- Determine the query domain by computing the vector space between query terms.
- Using the VSM to compute the query vector with documents vector, and go directly into the documents that have the same domain.
- Match the documents that related to the user query using Cosine similarity.
- Rank the related documents.

A. Keyword Extraction

Keywords extraction is a method to discover the terminologies that represent the document, and text contents [8]. It is very important to access to the main topics in the document, where it make the searching process so efficient. Keywords extraction also provides the target domain of document that helps the system to extract the relationships between concepts, and knowledge [10], [12].

B. Ontology

In this approach, we will use the Arabic ontology to access the relation between words, to candidate the keywords. Ontology includes several concepts which related with each other in class hierarchies. It concerns to determine the relevant concepts in an ontology, and semantic relations between of them. The ontology represents as a base infrastructure of knowledge. So the most researches of knowledge representation, semantic web concerned with ontologies. It also helps to determine the domain of the knowledge [6].

C. Preprocessing Phase

This phase is an important process in both, document, and query. The input of this phase is text of modern standard language of Arabic Newswire. It used to reduce the noise in the texts, through remove irrelevant or not important words such as stop words, prepositions, punctuation marks, digits from Arabic texts. As result, replace some Arabic letters into other letters to be more understandable and readable by computer. These subtasks can be summarized as follows:

1) *Text normalization*: This process is applying on several natural language texts. It represents a task to transfer the inconsistency text to be more consistency. In the Arabic language was used normalization to remove the diacritics marks, and normalize the other specific characters.

2) *Tokenization*: Tokenization is a process to divide the plain text into tokens to remove the noise from the text. After that sent it into the morphological analyzer to continue the processing [4].

3) *Stop words removal*: This process is to remove the frequent Arabic words that insignificant words or don't carry important meaning.

D. Matching

The document and query represent as vectors to determine the domain of them by using SVM. This process has been done by computing the similarity between the keywords vectors, and terms in the same document, and query. As a result, to that, the document in the same query domain, and related to the query are ranked and retrieved to the user.

V. DISCUSSION

Suppose having three document which taken from Jordan newswire named "Sarayanews", and its URL is "sarayanews.com" as shown in Table I. Each document has different text as the following example in table:

TABLE I. SIMPLE EXAMPLE TO ELABORATE THE PROPOSED METHOD

Doc #	Document content	Translate to English	The candidate keywords	Document Domain/ Category
Doc 1	يعترض المشرفون على شؤون كرة القدم في ألمانيا على فكرة زيادة عدد الدول المشاركة في المونديال ويرون أن ذلك سيؤثر على جودة البطولة. بالمقابل تزداد دول أخرى هذه الزيادة التي كان رئيس الفيفا قد وعد بها خلال حملته الانتخابية	German football supervisors refuse to increase the number of countries participating in the World Cup and see that this will affect the quality of the tournament. Nevertheless, other countries support the increase that the FIFA president had promised during his campaign	كرة القدم, المونديال, البطولة, الفيفا Football, World Cup, FIFA.	Sport
Doc 2	فاجأت أسرة البرنامج أعضاء لجنة التحكيم بعرض صور لهم خلال مرحلة الطفولة، غير أن اللافت أنه لم تُعرض صورة الفنانة الإماراتية أحلام على المسرح كسائر أعضاء اللجنة	The family of the program surprised the members of the arbitration by presenting their pictures during childhood, but it is remarkable that the image of Emirati artist Ahlam was not shown on stage like the other members of the committee	البرنامج, لجنة التحكيم, أعضاء Program, Committee, Members, judgements	Art
Doc 3	فقد مجلس النواب نصابه القانوني بعد أقل من نصف ساعة على بدء الجلسة الصباحية اليوم الاثنين وجاء فقدان النصاب دون رفع الجلسة المخصصة لمناقشة مشروع قانون الموازنة العامة وموازنات الوحدات الحكومية	The Parliament lost its quorum since less than half an hour after the start of the morning session on Monday and the loss of the quorum without lifting the meeting to discuss the bills of the budget and budgets of government units	مجلس النواب, نصاب Load, Parliament	Political

TABLE II. SAMPLE OF 3-QUERIES TO FIND RELATED DOCUMENTS

Query #	In the Arabic language	Translate to English
Q1	موعد بث برنامج ارب ايدول	Broadcast time of program "Arab Idol"
Q2	موازنة المملكة الاردنية الهاشمية 2017	Budget plan of Jordan 2017
Q3	عدد افرقة كأس العالم	Number of World cup teams

TABLE III. EXPECTED OUTPUT

Query #	Terms	Translate to English	Query Domain
Q1	بث, برنامج, موعد اربو ايدول	Broadcast date, program, Arab Idol	Art
Q2	موازنة, المملكة الاردنية, الهاشمية	Budget, Kingdom Jordan Hashemite	Political
Q3	عدد, افرقة, كأس العالم	Number, teams, cup, world	Sport

Suppose, we have three queries as shown in Table II:

As can be seen in above Table II. When we apply the all phases, the proposed model over these queries, the analysis result will be as shown in Table III.

The proposed system will be able to search about the related document in the same domain of query directly. This work is about Arabic information retrieval analysis through text classification application to access the information, or document that need within a certain domain. Domain identification is so benefited for both search engine, and information retrieval systems to retrieve the target information.

VI. CONCLUSION

Due to the amount of the available documents and texts in websites which is a challenge for information retrieval researchers' to minimize the required time to retrieve the documents to enhance the degree of performance and accuracy in the retrieval, requires more efforts. So, to solve these challenges, the information retrieval systems, accuracy and recall measurement should be enhanced. One of these suggests a solution to improve the efficiency of the Arabic information retrieval system is using text classification. Text classification helps information retrieval systems to access the target information domain. This paper showed the importance of text classification to improve the information retrieval from different resources. It aims to enhance the performance of information retrieval systems.

ACKNOWLEDGMENT

We would like to thank University Malaysia Terengganu (UMT) for providing us with a good scientific environment to produce this simple work

REFERENCES

- [1] Ababneh, J., Almomani, O., Hadi, W., El-Omari, N. K. T., & Al-Ibrahim, A. (2014). Vector space models to classify Arabic text. *International Journal of Computer Trends and Technology (IJCTT)*, 7(4), 219-223.
- [2] Al-Kabi, M. N., & Al-Sinjalawi, S. I. (2007). A comparative study of the efficiency of different measures to classify Arabic text. *University of Sharjah Journal of Pure and Applied Sciences*, 4(2), 13-26.
- [3] Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *Int. Arab J. e-Technol.*, 2(2), 124-128.
- [4] Attia, M. A. (2007). Arabic tokenization system. Paper presented at the Proceedings of the 2007 workshop on computational approaches to Semitic languages: Common issues and resources.
- [5] Belkredim, F. Z., El-Sebai, A., & Bouali, U. H. B. (2009). An ontology-based formalism for the Arabic language using verbs and their derivatives. *Communications of the IBIMA*, 11(5), 44-52.
- [6] Brewster, C., & O'Hara, K. (2007). Knowledge representation with ontologies: Present challenges—Future possibilities. *International Journal of Human-Computer Studies*, 65(7), 563-568.
- [7] El Kourdi, M., Bensaid, A., & Rachidi, T.-e. (2004). Automatic Arabic document categorization based on the Naïve Bayes algorithm. Paper presented at the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.
- [8] Elghannam, F., & El-Shishtawy, T. (2015). Keyphrase based Evaluation of Automatic Text Summarization. *arXiv preprint arXiv:1505.06228*.
- [9] Froud, H., Lachkar, A., & Ouatik, S. A. (2013). Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. *arXiv preprint arXiv:1302.1612*.
- [10] Harith, A., Kim, S., Millard, D. E., Weal, M., Hall, W., Lewis, P., & Shadbolt, N. (2003). Automatic ontology-based knowledge extraction and tailored biography generation from the web. *IEEE Intelligent Systems*, 18(1), 14-21.
- [11] Harrag, F., & Al-Qawasmah, E. (2010). Improving Arabic Text Categorization Using Neural Network with SVD. *JDIM*, 8(4), 233-239.
- [12] Hasan, K. S., & Ng, V. (2014). Automatic Keyphrase Extraction: A Survey of the State of the Art. Paper presented at the ACL (1).
- [13] Moh'd A Mesleh, A. (2007). Chi-square feature extraction based SVMs Arabic language text categorization system. *Journal of Computer Science*, 3(6), 430-435.
- [14] Sawaf, H., Zaplo, J., & Ney, H. (2001). Statistical classification methods for Arabic news articles. *Natural Language Processing in ACL2001*, Toulouse, France.
- [15] Thabtah, F., Eljimini, M., Zamzeer, M., & Hadi, W. (2009). Naïve Bayesian-based on chi-square to categorize Arabic data. Paper presented at the proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt.

AUTHOR'S PROFILE

Mr. Mohammad Khaled A. Al-Maghasbeh is an PhD researcher of computer science in School of Informatics and Applied Mathematics, at Universiti Malaysia Terengganu (UMT)- Malaysia. He obtained his M.Sc. in Computer Science from Al-Balqa'a Applied University, Jordan in 2013 with dissertation titled "Agent-Based Data mining for proteins prediction and classification". He got his B.Sc. in software Engineering from Al-Hussein Bin Talal University (AHU) in 2009. His interest and active researches are in Data mining, Big-data, Artificial Intelligence, Information Retrieval, and Natural Language Processing (NLP).

Features and Potential Security Challenges for IoT Enabled Devices in Smart City Environment

Dr. Gasim Alandjani
CSE-ICT Department
Yanbu University College
Yanbu Alsinayah, Kingdom of Saudi Arabia

Abstract—Introduction of Internet of Things in our lives have brought drastic changes in the social norms, working habits, ways of completing tasks and planning for future. Data about our interactions with everyday objects can be effectively transmitted to their destinations with many communicating tags that also often provide specific location information. The risk of potential eavesdropping is always a major concern of data owners. Since Internet of Things is primarily responsible for carrying data of smart objects which are mostly connected over wireless technologies, securing of information carried by these wireless links to safeguard the private information is of utmost importance. Cryptographic techniques to cypher data carried by the IoT networks is one possibility which is not feasible due to the lack of sufficient computing resources at the sensor end of IoT devices. In this paper, we discuss various security issues that haunt the secure IoT deployments and propose a layered solution model that prevents breach of security during transmission of data.

Keywords—IoT; privacy; smart city; smart society; actuators; sensors; industrial 4.0;5G

I. INTRODUCTION

Technology is taking us to the next level for providing end users with state of the art services by using latest cutting-edge technologies. As far as security is concerned for all these latest technology we can take example of Internet which is still not secure, so same is the case with other technologies and eventually there is no expectations that IoT will be secure. However, with the passage of time security is constantly evolving to meet new challenges and also addressing the old ones that we've faced in past, and we'll see them again, with IoT and succeeding associated technologies.

Leading companies have stopped development for old technologies and shifted development for latest cutting-edge technologies, recent example is Intel who drops plans to develop spectre microcode for ancient chips.

New manufacturing processes generally result in faster and more efficient processors, and time is not far when this gap will close, thus providing developers with enough processing power in these devices to implement enhanced and better security features.

According to research firms International Data Corporation (IDC) and Gartner, IoT will grow to technology to such advance level that which will change layout and processing requirements which are available in current format of data centers. Gartner predicts by 2020 IoT market will have

26 billion devices will have IoT enabled sensors, which eventually will be creating huge opportunities for hardware manufacturers, data centers users, and developers. IDC also expects huge investment in IoT industry with "billions of devices and trillions of dollars" by the end of the decade and resulting with the following potential challenges.

- Enterprise: Security issues could pose safety risks.
- Security: Increased automation and digitization creates new security apprehensions.
- Data: Tons of data will be generated, both for big data and personal data.
- Consumer Privacy: Potential of privacy breaches.
- Data Centre Network: WAN links are optimized for human interface applications, IoT is expected to dramatically change patterns by transmitting data automatically
- Server Technologies: More investment in servers will be necessary.
- Storage Management: Industry needs to figure out a cost-effective way to deal with tons of data generated by these IoT enabled sensors.

As technology is getting smarter there is great increase in popularity of tinny technological gadgets including Smart wristbands, toasters and dog collars which aren't a huge concern from a security perspective, due to low cost and there is lack of processing power in these devices which is another security problem, as most advanced encryption techniques simply wouldn't work very well, on the other hand if more processing and storage capacity is added to these devices which will eventually increase their cost and will throw than out of the competition of these popular devices. In a Survey HP revealed that 70% of IoT devices are vulnerable to attackers.

Here is a list of points to consider that can help in improving security.

- Security emphasis from day first
- Lifecycle, future-proofing, updates
- Consideration for Access control and device authentication
- Never underestimate power of hackers

- Well Prepare for possible security weak points and their solutions.

Based on usage, network location and processing power of IoT devices the level of threat varies from device to device and there are uncountable concerns to consider while using them for domestic purpose end users should have sufficient knowledge about all these threats before they start using these devices at homes and offices for personal use.

Users should be ready for potential security breaches. As they are inevitable, it can happen to you or someone else. Make sure that you should always have a solution for any possible security breach for maximum security of data and interpreting compromised data useless without breaking your IoT infrastructure(most the time It infrastructure in offices is and will be more secure as compared to normal users who will be using these devices for personal use at their homes.

If they are interested in expanding services through the IoT then they must keep consumer choice and preference while deciding which capabilities they would deliver on a smartphone versus a smart watch. Similarly, a Mobile App Development Service Provider should use the same lens while developing applications for those connected devices.

While talking to user end services in smart city, which is offering a vast range of device automation and management at user side. The major security issues in IOT field are confidentiality, authentication, access control, trust, mobile security, privacy, policy enforcement and secure middleware.

II. RELATED WORK

Security in IOT is very interesting topic these days. Many projects are started in this context. One of the projects is Butler which is European Union FP7 projects. It provides secure context-aware and location aware services to assist smart home, city, hospitals and business domains. [1] describes about Iot applications and their interaction with each other based on different nature of hardware interfaces different devices are not able to communicate properly so is the case with different types of applications which have been designed but still there are some missing dots that need to be filled to get maximum from these IoT based application which have been used to provide different services.[2] describes the Hydra project develops middleware for network embedded systems based on service oriented architecture. This project deals with security issues and trust issues among distributed components of middleware. The role of middleware is to incorporate among heterogeneous devices using different technologies. [3] Describes basic principles with methodology of experiment which will be bridging social network interactions and sensor measurements. Its aim is to exploit the smartSantander for sensor measurement and communication to the public. And also to analyze and summarized sensor reporting and development of collective aware applications. [4] Describes uTRUSTit project which is usable trust in Internet of things. It offers the trust feedback toolkit in order to enhance user security. [5] Describes for consideration of a particular city as smarter one based on different practices. It has used a set of multidimensional components as a core factor for smart city and successful delivery of its services. It also offers strategic

principles aligning to technology, people, and institutions of smart city and it further goes to show human learnings based on these facts. [6, 7] describe that most studies on practices of smart city address issues of technological infrastructure and associated enabling technologies. The focus on state of the art infrastructure will help technology, accessibility and availability of systems. [8] describes the iCore project which is large management system for IOT in ecosystem. It consists of following components VO virtual objects, composite virtual objects and real world objects. In USA there are also many IOT based projects. One of the projects is proposed by DARPA which is High Assurance Cyber Military System (HACMS), it assures that military vehicles, equipment and drones cannot be hacked from outside. Roseline is another project which is issued by NSF. Its work is to enhance the robustness in cyber-physical systems. Furthermore NSF projects are: XIA-NP (Development-Driven Evaluation and Evolution of the expressive Internet Architecture) which describes the diversity in network models, NDN-NP(Named Data Networking-Next Phase addresses the technical challenges like routing, scalability, fast forwarding, trust models and privacy. NEBULA provides architecture for cloud computing, and Mobility next Phase describes general mobile delay tolerant. These projects will explore novel network architecture and protocols. Further projects are included by National Basic Research Program focuses on the Security Protection among different entities of IOT.FIRE (Future Internet Research and Experimentation) is a project of Europe, China and Korea which is realization of different IOT technologies in different areas. EU-JAPAN cooperate for developing global standards and seamless communication.[9] describe a cloud model for provision of efficient services to the end users without compromising their personal security which using cloud any community cloud, it further describe different available solutions based on different types of clouds e.g public cloud, private cloud and community cloud. [10] describes a triple-helix model which enable to study the knowledge base of an urban economy for local community support regarding evolution of key components of innovation system, it further claims that cities can be considered as the intellectual capital of universities, the wealth creation for industries and democratic government for civil society interaction of these three densities generate dynamic spaces where knowledge can be used for bootstrap as technology for regional systems.[11] describes the conceptual scene for city e-governance, with a major focus on creation of cooperative digital environments to enable local competitiveness and prosperity through knowledge networks and partnerships, it further showed results of a very detailed survey study in which was conducted in twelve European cities. [12]describes smart infrastructure framework development supported by survey regarding accuracy for position of any devices which have been used for providing services to the inhabitants of smart society, it also discusses main advantages of proposed architecture with measureable and non-measureable benefits. [13] describes a smart innovation ecosystem characteristics which clarify the assembly of all smart city concepts into green , open, instrumented, integrated and intelligent layers which further compose a planning frame work which is called smart city reference model based on different shapes and sizes

of cities. This model can be used to for smart policy paradigms and encirclement the green, broadband and urban economies. [14] Explains the industrial 4.0 where human will be replaced by AI based robots which can be controlled by augmented reality based on needs and typical routine requirements of work flow and in case of emergency to control the delicate processes and critical situations. [15] Highlights the issues that can be a reason in increase of multidimensional challenges for both (city residents and administration) entities of smart cities and further proposed a conceptual framework of cloud based architecture context aware smart services for inhabitants of smart cities. [16] it discusses about some standard navigation system which help to create a navigation model which can be used to find location of service providing devices installed on different locations of smart city. [17] described about future trends which will be going to create a community where classes will be defined based on community services for different classes and it will further create an environment of a jail-less community where there is be no conventional jail for criminals rather they will be deprived of some services and they have to inform local police before leaving premises of smart city.

There is enormous pressure on the city management to provide sustainable services and facilities to the growing cities paving the way to launch smart city initiatives by the government, public and private sector. IoT has also gained importance in smart city development. IoT facilitate people and things to connect with each other at anytime, anyplace, with anyone by using any network to access their required services. Smart city concept revolves around six fundamentals namely, smart people, smart governance, smart economy, smart mobility, smart living and smart environment. Smart City and IoT are evolving together to achieve the same goals. IoT heavily relies on cloud services for data consolidation, big data analytics, reporting and web front-end etc. Everything as a service (XaaS) is the concept offered by cloud to offer different levels of services as per the requirements of the end users or devices. The basic idea behind cloud computing and storage is to concentrate resources such as hardware and software into geographically diverse locations and offer those resources as service to large number of consumers who are located in many different geographical locations. There are three well defined levels of cloud services i.e. Platform-as-a-Service (PaaS), Infrastructure-as-a-Service (IaaS), and Software-as-a-Service (SaaS). Based on these models sensing as a Service model is designed to address solutions for IoT and challenges in Smart City. It consists of four Conceptual Layers as mentioned subsequently:

III. MAJOR SECURITY CHALLENGES

There a numerous security challenges for IoT devices with current available infrastructure as of now and providing connectivity in smart city environment.

A. Technical Sophistication Gaps

A multifaceted system of connected devices opens many new attack vectors, even at individual level if each device is secure while not connected to network. Since a system's most vulnerable point decides its overall security level, a

comprehensive, end-to-end approach is required to secure it. Which is very difficult to develop.

B. Absence or Immaturity of Standards

The IoT lacks well-established predominant standards [17] that describe about different components of technology should interact. Some segments, such as industrials, still rely on a small set of proprietary, incompatible technology standards issued by the major players, as they have done for many years. In other segments, such as automotive or smart buildings, standards are basic. Development of end-to-end security solutions in absence of common standards will be difficult task for IoT device manufacturers.

C. Consideration of IoT as Commodity

With all new productive gadgets of IoT majority of customers still consider it more as commodity rather considering it a mature product that's the main reason they don't think to go for security of these devices.

D. Challenges for Manufacturers

Most of the semiconductor manufacturing companies are currently struggling a lot to embed security features during manufacture process as it result in high cost and difficult to meet market cost effective demands. One side role f IoT in smart buildings is expected to increase by 40%, On the other hand IoT security breaches are rising in residential applications. This security trend may vary at user end based on their usage behavior e.g. some users might update firmware continuously on the other side some might not be updating them which will eventually become a potential security risk for these devices.

There is a great need to propose a sensor network model which follow the layered approach and get data in a systematic way from different sensors according to requirement for communication.

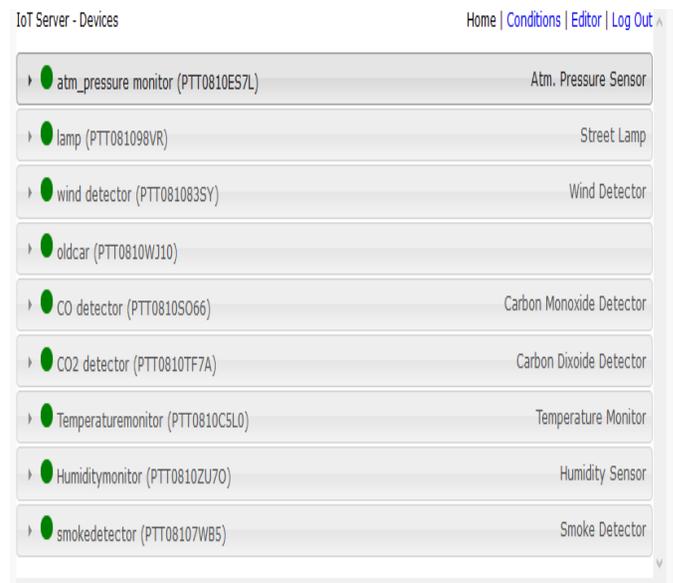


Fig. 1. Data Collection through Different Sensors.

This data can be collected by using different low computing devices including smart phones. Below figures are

depicting different types of data collected by different sensors and then further plotted them against different values which smart city users will be using collecting data of smart environment to show it from different aspects.

Fig. 1 is showing data of different sensors which we have calibrated by use of Cisco Packet Tracer software, where we can add different sensors and read output for different values in any given scenario before its physical deployment.

Fig. 2 is showing different levels of atmospheric pressure at different intervals of time, these values have been taken in normal situation, in case if some unauthorized person manage to get access to this system with intention to alter it for some specific goals results could catastrophic as people might be relying on output values provided by these sensors. In case they get some wrong results at some critical time which might generate great loss.

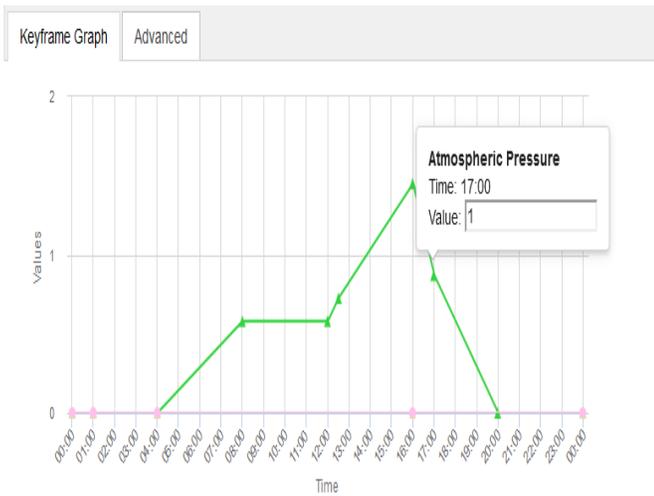


Fig. 2. Atmospheric Pressure.

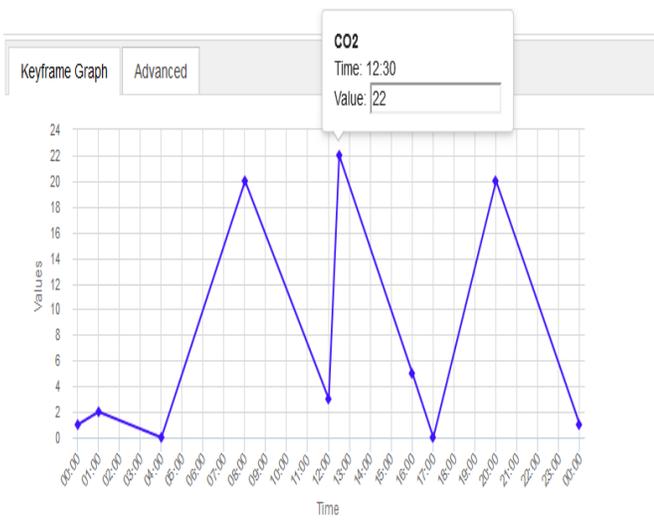


Fig. 3. Carbon Dioxide Levels.

Fig. 3 is showing different levels of Carbon Dioxide at different intervals of time, these values have been taken in normal situation, in case if some unauthorized person manage

to get access to this system with intention to alter it for some specific goals results could catastrophic as people might be relying on output values provided by these sensors. In case they get some wrong results at some critical time which might generate great loss if there is extra ordinary increase in CO2.

Fig. 4 is showing different levels of ambient temperature variation at different intervals of time throughout the whole day, these values have been taken in normal situation, in case if some unauthorized person manage to get access to this system with intention to alter it for some specific goals results could catastrophic as people might be relying on output values provided by these sensors. In case they get some wrong results at some critical time which might generate great loss if there is extra ordinary change in environmental temperature and it might further make is critical for industry especial in the presence of industrial 4.0 if it goes unnoticed due to fake readings presented through any compromised IoT monitoring system.

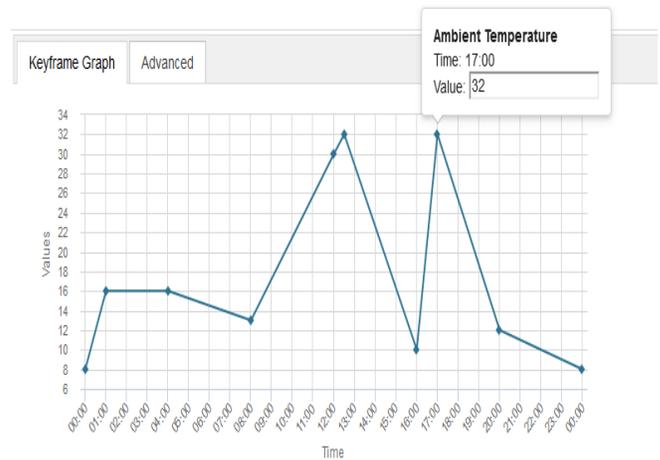


Fig. 4. Ambient Temperature Variations throughout a Day.

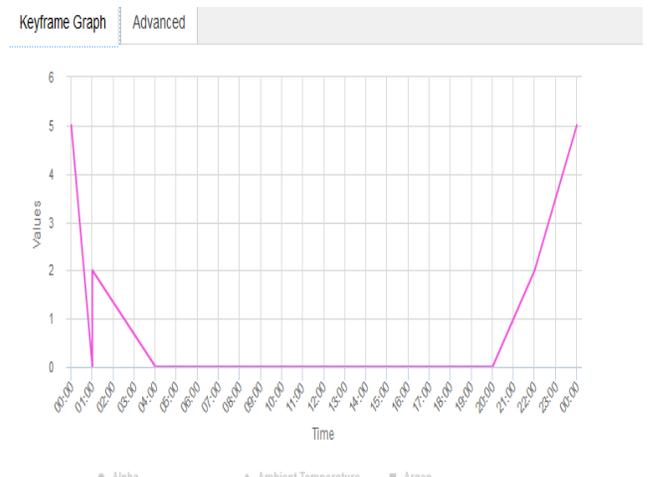


Fig. 5. Humidity Level throughout a Day.

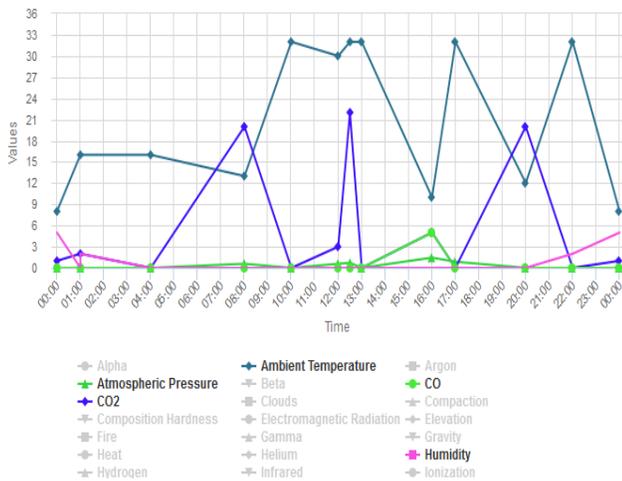


Fig. 6. Sensors Data Graphical View.

Fig. 5 is showing different levels of humidity at different intervals of time, these values have been taken in normal situation, in case if some unauthorized person manage to get access to this system with intention to alter it for some specific goals results could catastrophic as people might be relying on output values provided by these sensors. In case they get some wrong results at some critical time which might generate great loss if there is some extra ordinary change in humidity.

Putting all these together, Fig. 6 is showing all data in a single graph for values of all these sensors. Showing them at a single graph help to read understand overall data trends for different sensors. In case they get some wrong results about any of the above sensors at some critical time which might generate great loss if there is extra ordinary change in environmental temperature, CO₂, Humidity, atmospheric pressure and humidity level. it might further make is critical for industry especial in the presence of industrial 4.0 if it goes unnoticed due to fake readings presented through any compromised IoT monitoring system.

IV. IOT SECURITY SERVICES

Following services are used by IoT devices in smart city environment sharing data through sensors.

A. Authentication

For authentication and confidentiality they have discussed work proposed by various researchers. One of the proposed method is custom encapsulation mechanism which includes encryption, signature and authentication. The two way security authentication scheme is also very popular. It uses Datagram Transport Layer Security (DTLS) protocol which is present between transport and application layers. It uses RSA which is designed for IPv6 low power wireless personal area network (6LoWPAN). It provides integrity, confidentiality and authenticity with affordable energy, end-to-end latency and memory overhead. The key management system has four major categories key pool framework, mathematical framework, negotiation framework and public key framework. Some of the KMS protocols are not suitable for IOT ,for example key pool framework has connectivity issue, mathematical evaluation needs optimization to construct data

structure, however some of the KMS are effective like Blom and polynomial schema whose computational overhead is low and use public key infrastructure (PKI). A transmission model which uses signature-encapsulation schemes and provides anonymity, attack-resistance and trustworthiness. This model utilizes object naming service (ONS) queries. Root ONS authenticates queries by local ONS via trusted authentication server (TAS) and prevents illegal ONS. Remote information server of things (R-TIS) wraps the information in encryption layer with the public key of routing node. The information is routed at every node until the local information server of things (L-TIS) receives plain text. However this method is weak in attack-resistance. Although above methods provide better security in terms of confidentiality and authentication but some questions are really answerable i.e. at which layer we should apply security mechanism, how to handle keys, which key distribution method will be useful, can we use previous authentication mechanism and how to apply end-to-end integrity to prevent malicious attacks. Some of the recent work to address such questions is authentication mechanism for IOT using lightweight encryption using XOR manipulation for anti-counterfeiting and privacy protection, for WSN user authentication and key agreement between users and remote sensors and another lightweight encryption mechanism called elliptic curve cryptography (ECC) for authentication and attribute based access control.

B. Access Control

Access means how different resources are provided to different users. Two terms are frequently used; data holders which are users and things while data collectors are sensors and service providers. In IOT data streams have to be processed and many queries are generated so enough data manipulation is needed. Every node is given a limited computational, storage capacity and single key. Other keys are manipulated so storage capacity is saved. The authentication system for emergency cases e.g. in case of accidents availability, name and location must be provided. Nile security architecture is also very popular which process data streams by frequent queries using cipher encryption and decryption keys. The authentication process for the outsourced data (in cloud computing). It involves authentication from the source and process queries for clients so data from authenticated sources are processed and clients get the right.

C. Trust

Trust concept is related to security and access control. The researchers have described how devices are heterogeneous, different users share friendship and belong to different community so malicious attacks are common. Self-promoting, good mouthing and bad mouthing are trust related attacks. The trust management protocol. It is distributed, encounter-based and activity based.it means that when two devices communicate with each other they perform trust based evaluation with each other. The evaluation parameters are honesty, cooperativeness and community interest. The reputation based trust mechanism for the IoT nodes to prevent malicious node and ensure communication for the trusted nodes only. They proposed a subjective model for P2P devices, in this model each node computes the trustworthiness of the neighbor node and ensures communication with only

trustworthy node. The secure ad-hoc networks it provides peer to peer communication and communities to surf web. It involves following parameters to analyze; physical proximity, fulfillment, consistency of answers, hierarchy on trusted chains, similar properties, common goals, availability and interactions. The phenomenon of fuzzy approach to trust based access control (FBTAC). Trust scores are calculated by factors like experience, knowledge and recommendations. It consists of three layers device, request and access. Device layer consists of all devices, request layer consists of all the recommendations and fuzzy results and access layer involves decision making. This fuzzy approach provides flexibility and scalability. It is easier in utility based decision making. Fuzzy approach based upon three layers; sensor, core and application layers. Sensor layer consist of physical devices like sensors and RFID, the core layer consist of access network and internet. Application layer consists of distributed networks (e.g. P2P, grid, cloud computing). Evaluation of trust management by fuzzy set theory and semantic based language on layered approach and layer attributes are history, risk and efficiency. In this model user can access to the IOT devices only if the security credentials are satisfied. A trust model by utilizing the location, identity and authentication history. There are three trust regions based on trust levels:

- i. High trust level
- ii. Medium trust level
- iii. Low trust level

In high region of trust no authentication is required only VID is used. In medium region users offer their PIN to login, in low region of trust different authentications are required like face identification, fingerprints and iris scan. The trustworthiness of nodes by their past behavior. It involves following steps;

- i. Gathering of information about the trustworthiness of neighboring nodes.
- ii. Set up collaborative service with neighboring nodes.
- iii. Learn about the previous operation and update.
- iv. Assign a quality recommendation score to each node.

Attack resistant model is proposed by researchers for distributed approach. It provides trust in self-organized nodes and attack resistance in distributed nodes. The WSN nodes and provides identity based network to the devices. It prevents attacks from the malicious nodes. The identity management systems for nodes which move from host to host so they need location and identification to separate from host addressing. Following techniques have been used, to achieve trust, so for social networking, fuzzy approach, identity based networking and cooperative approach. Following issues are still open in Trust management.

- i. Introduction of semantic based language for the negotiation of trust.
- ii. Proper identity management system.
- iii. Development of trust management system for data stream control.

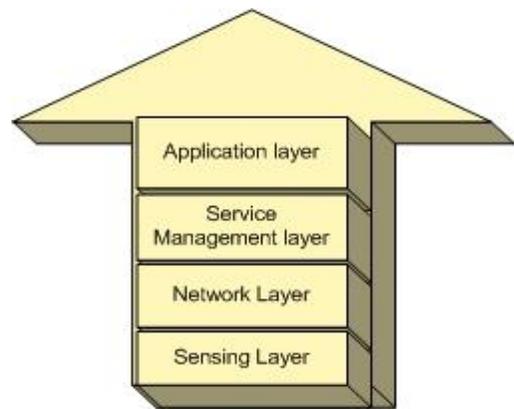


Fig. 7. Layered Architecture of IoT.

D. Mobile Security in IOT

Mobile nodes move from one cluster to another so rapid identification, authentication and privacy protection is required. The ad-hoc network protocol when nodes move from one cluster to other. It uses request messages and answer message for identification, authentication and privacy protection. This process has less overhead, more security and protection. HIMALIS (Heterogeneity Inclusion and Mobility Adaptation through Locator ID separator).it proposes secure and scalable mobility management. It provides secure inter domain authentication, secure location update and binding transfer for mobility process. RFID system based upon Electronic Product Code (EPC). It explains the mobile threats of RFID nodes. It guarantees security and efficiency. The security of tag and readers are also very important aspects. It also explains tag corruption, reader corruption, multiple readers and mutual key exchange protocol. The location security issues in mobile nodes. It pays attention to special location issues in android, iPhone and windows network platform. the secure handshake between mobile nodes in intelligent system is also a prime concern. Mobile node verifies the legitimate sensor node over an insecure channel via negotiation of handshake protocol. The mobile solution for healthcare services. It provides security and privacy mechanism for the security of the patients. The RFID tag identification and IOT infrastructure is combined. Efficient and secure mobile intrusion detection system for business applications using human centric computing is in placee. The mobile information collection through IOT gateway via smart devices. Quantum Lifecycle Management System messaging standard to provide two way communications between firewalls is also very interesting idea for security. Mobile Sensor Data Processing Engine (MOSDEN) is another technique. It collects and processes sensor data without programming efforts .it uses plug-in based IOT platforms for mobile devices. Other techniques are discussed by different researchers like video dissemination for IOT devices, interaction of smart things via Bluetooth and use NFC via mobile devices via Web of Things.

1) *Proposed Solutions:* As there is no complete solution related to internet security, same is the case with IoT up till now there is no single major solution regarding security and privacy of IoT devices, infect IoT devices are more prone to

hackers and attackers due to the fact that IoT communication is primarily a sensor based communication which further make it more independent when devices start communicating to other devices without waiting for permission from human or without interference of human. Another thing make them more prone to attacks because most of the these devices are standalone and if their firmware is not updated at regular intervals it will increase chances of attacks so to avoid this there should be regular firmware updates on all these standalone devices as suggested by manufacturers. Apart from all above suggestion the most reliable approach for IoT security which is suggested by most of the researchers is to divide security into different levels and it will help to stop attackers directly accessing devices most commonly known technique is called the layered approach, Fig. 7 is showing a layered model suggested to avoid direct security attacks on IoT devices.

To elaborate the notion of smart world and its smart components there are many research communities focused on IoT, mobile computing, wireless sensor networks and cyber-physical system. Research in these areas relies on machine learning, real-time computing, security, privacy and signal processing. Fig. 8 is showing authentication procedure through Sequence diagram for QR Based authentication. Our living style and working habits will be changed significantly with the inclusion of these new technological trends. IoT in many different angles cover including architecture, massive scaling, dealing with big data, focusing on security, privacy.

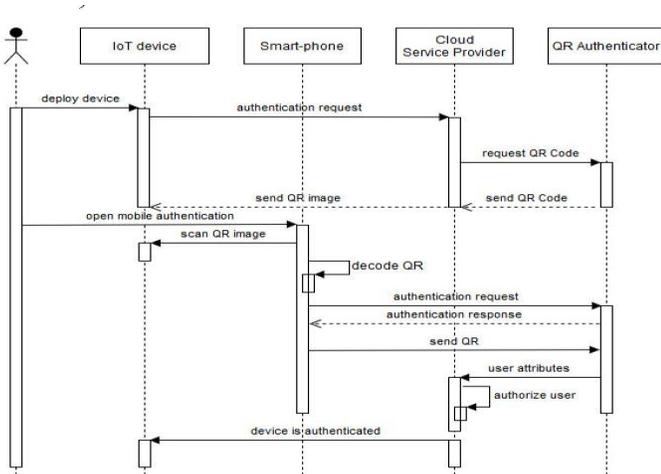


Fig. 8. Sequence Diagram for QR based Authentication.

- **Massive Scaling:** there is a prediction for trillions of devices on the Internet which is going to be a big challenge to deal with security and privacy aspects on such a large scale.
- **Architecture and Dependences:** connectivity such a massive scale devices on internet require a well-defined architecture that allows communication, control, and useable apps.
- **Creating Knowledge and Big Data:** IoT require huge amount of raw data to be collected so there is need to

develop technique to convert this data into information. Data mining techniques are used to extract knowledge from sensors data. Another main challenge while extracting knowledge is making decisions while minimizing the false positive and false negative and guarantee safety.

- **Robustness:** In IoT applications sensing, actuations and communication is needed. Each node must be aware of other node's location and synchronized clock. Clock drifts because nodes to have different times resulting in application failure. So for the collections of solutions to create robust systems.
- **Openness:** It means that system is continuously changing and devices have to communicate with each other in this system efficiently. Many sensors and actuators use control and feedback mechanism via controllers.
- **Security:** The fundamental problem in IoT is protection from security attacks. Security attacks create problem due to limited capacity of devices. There must preemptive security measures to protect from these attacks.
- **Privacy:** To solve the privacy problems of IoT the privacy policies of the each system must be specified and enforced accordingly.
- **Humans in the Loop:** Many IoT applications involve humans in the process. Although humans in the loop have many advantages but modeling human behavior is difficult due to physiological, psychological and behavioral aspects

E. Proposed Layered Security Approach

Fig. 9 shows proposed block diagram for data collection and its security during transmission of IoT data.

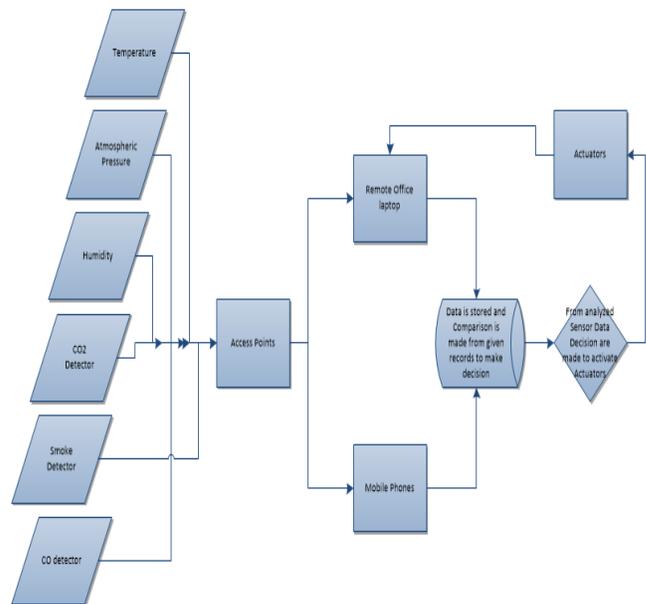


Fig. 9. Proposed Sensors Network Model.

Which will be taking sensors for temperature, atmospheric pressure, humidity, CO2 detector, Smoke detector, CO detectors, these sensors will be transmitting data to any of the near available access points which will be transmitted that data to remote laptop/desktop or on mobile phone. This data will be further stored in knowledgebase as record for future references and decision making and this analyzed data will be used for decision making to enable actuators which will be further sending updated data to remote computers. Based on nature of communication devices the IoT provides more number of vulnerable points for security breaches to occur, it is very critical to have multi-layers of security. This is because if one of the layers is breached then you must have other mechanisms to fall back on.

V. DISCUSSION AND FUTURE WORK

This paper discusses different security issues which residents of smart cities are facing and it also provide solution for all these challenges. Future work related to these security issues can be done by registering all the end-user devices in a central data base and all the data stored should be in encrypted form which at one end can increase retrieval time but at the other end it will make sure security of data for all the users who will be using different services provided by smart city administration.

ACKNOWLEDGMENT

I really appreciate cooperation from management of Royal Commission Yanbu and colleagues of University College for facilitating me in setting up lab work.

REFERENCES

- [1] Nasser H. Abosaq, Gasim Alandjani, Shahbaz Pervez. "IoT Services Impact as a Driving Force on Future Technologies by Addressing Missing Dots". International Journal of Internet of Things and Web Services, 1, 31-37, April-2016.
- [2] M. Jahn, Ferry Pramudianto, A.-A. Al-Akkad, "Hydra middleware for developing pervasive systems: A case study in the e-health domain", January 2009.
- [3] Vakali, A., Angelis, L., & Giatsoglou, M. (2013). Sensors talk and humans sense towards a reciprocal collective awareness smart city framework. IEEE International Conference on Communications Workshops (ICC).
- [4] Kourtiti, K. et al. (2013). 11 An advanced triple helix network framework for smart cities performance. Smart Cities: Governing, Modelling and Analysing the Transition 196.
- [5] Pardo, T., Taewoo, N. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. Proceedings of the 12th Annual International Conference on Digital Government Research (pp. 282–291). ACM, New York.

- [6] Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., PichlerMilanoviü, N., & Meijers, E. (2007). Smart Cities: Ranking of European Medium-Sized Cities. Vienna, Austria: Centre of Regional Science (SRF), Vienna University of Technology. Available at http://www.smartcities.eu/download/smart_cities_final_report.pdf.
- [7] Giffinger, R., & Gudrun, H. (2010). Smart cities ranking: An effective instrument for the positioning of cities? ACE: Architecture, City and Environment, 4(12), 7-25. Available at http://upcommons.upc.edu/revistes/bitstream/2099/8550/7/A_CE_12_SA_10.pdf.
- [8] Smart cities: ranking of European medium-sized cities. Centre of Regional Science (SRF), Vienna University of Technology, Vienna, Austria, from http://www.smart-cities.eu/download/smart_cities_final_report.pdf
- [9] Shahbaz Pervez, Faheem Babar, Gasim Alandjani, "An Efficient Cloud Model with integrated Services by addressing Major Security Challenges.", Journal of World Scientific Engineering Assembly and Society Transactions on Computers Print ISSN: 1109-2750, E-ISSN: 2224-2872.
- [10] Leydesdorff, L., & Deakin, M. (2011). The triple-helix model of smart cities: a neo-evolutionary perspective. Journal of Urban Technology, 18(2), 53–63.
- [11] Paskaleva, K. A. (2009). Enabling the smart city: the progress of city e-governance in Europe. International Journal of Innovation and Regional Development, 1(4), 405–422.
- [12] Al-Hader, M., & Rodzi, A. (2009). The smart city infrastructure development and monitoring. Theoretical & Empirical Researches in Urban Management, 2, 11.
- [13] Zygiaris, S. (2013). Smart city reference model: assisting planners to conceptualize the building of smart city innovation ecosystems. Journal of the Knowledge Economy, 4(2), 217–231.
- [14] Industry 4.0: the fourth industrial revolution – guide to industry 4.0 <http://www.i-scoop.eu/industry-4-0/>
- [15] Z. Khan, S. Kiani, K. Soomro, "A Framework for Cloud-based Context-Aware Information Services for Citizens in Smart Cities", Journal of Cloud Computing: Advances, Systems and Applications, vol. 3, No. 1, pp. 14, 2014.
- [16] M Handte et. Al (2016), "An Internet-of-Things Enabled Connected Navigation System for Urban Bus Riders", IEEE Internet of Things Journal, Volume 3, Issue 5.
- [17] Shahbaz Pervez, Nasser Abosaq, Gasim Alandjani, Adeel Akram, "Internet of Things (IoT) as beginning for Jail-Less Community in Smart Society", "IEEE International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing 28-29 January 2018 at Tamil Nado India.



AUTHOR'S PROFILE

Gasim Alandjani received his PhD Computer Engineering degree from New Mexico State University (USA), He has 27 years' experience of teaching and research including management experience as Dean, Makkah College of Technology-2003-2009, Deputy Managing Director of Yanbu Industrial College 2010-2012, managing Director of Yanbu Industrial College 2012-2013. Currently, he is working as senior faculty Member in ICT Department at Yanbu University College Royal Commission Yanbu, Kingdom of Saudi Arabia.

Comparative Study of PMSG Controllers for Variable Wind Turbine Power Optimization

Asma Hammami, Imen Saidi, Dhaou Soudani

Automatic Research Laboratory, L.A.R.A, National Engineering School of Tunis, ENIT
University of Tunis El Manar
Tunis, Tunisia

Abstract—With a large increase in wind power generation, the direct driven Permanent Magnet Synchronous Generator is the most promising technology for variable speed operation and it also fulfills the grid requirements with high efficiency. This paper studies and compares conventional based on PI controller and proposed control technique for a direct driven PMSG wind turbine. The generator model is established in the Park synchronous rotating d-q reference frame. To achieve maximum power capture, the aeroturbine is controlled through Maximum Power Point Tracking (MPPT) while the PMSG control is treated through field orientation where the two currents control loops are regulated. A proposed direct-current based d-q vector control design is designed by the integration of the Internal Model Controller. Then an optimal control is developed for integrated control of PMSG power optimization and Voltage Source Converter control. The design system was done using *SimWindFarm* Matlab/Simulink toolbox to evaluate the performance of conventional and proposed technique control of PMSG wind turbine. The analysis, simulation results prove the effectiveness and robustness of the proposed control strategy.

Keywords—Wind turbine; internal model control; PI controller Permanent Magnet Synchronous Generator (PMSG); vector control

I. INTRODUCTION

Recently, wind energy system has been treated as an important renewable energy source which had higher potential to generate power where grids are not feasible. The wind generation systems have gained tremendous attention over fossil fuel and nuclear power generation due to the high cost and environmental clean [1]. At present the variable speed wind turbine is considered the most attractive solution to distribute power generation systems. Mainly four types of generators are used in wind power system: Squirrel cage induction generator – double fed induction generator – wound rotor synchronous generator – permanent magnet synchronous generator (PMSG). Considerable research has been devoted to the choice of PMSG for variable speed generation system. It has high efficiency, is connected directly to the turbine without gearbox and has full controllability of the system for maximum wind power extraction [2].

However, the performance of PMSG depends on the control strategy. Traditionally, PMSG with full scale PWM converter is controlled through the conventional decoupled d-q vector control. The overall problem that occurs in this method is the calculation for determination of controller parameters and the robustness performance. Most studies have used the

adaptive control scheme as a robust method of control strategy while others use the artificial intelligence techniques. These structures are required for exact mathematical identification of controller parameters.

The Internal Model Control method was observed by Gracia and Morari [3], [4], and was later improved under intensive research and development. This design provides high performance dynamic characteristics. This structure covers an internal model of the plant and an internal model controller. In order to improve the disturbance rejection, a modified IMC is designed with an additional filter [5]. It provides good abilities of control system performance particularity for the stability and robustness issues.

This paper proposes a comparative study between conventional vector control with PI action controller and proposed IMC controller design applied for the purpose of improving the control effectiveness and overall performance of PMSG system.

This paper is structured as follows: In section 2 a mathematical model of the wind turbine with PMSG system is described. Section 3 deals with the PMSG control including the IMC proposed control is developed in section 4. The designed and traditional controls are compared and the validation results using *SimWindFarm* Matlab/Simulink toolbox are shown in section 5.

II. WIND TURBINE MODELING

A. Aeroturbine Modeling

The aerodynamic blades allow the conversion of the kinetic energy of the wind profile into mechanical energy to the generator.

Therefore the aerodynamic torque T_a is given by [6]:

$$T_a = \frac{1}{2\omega_r} \rho A V_w^3 C_p(\lambda, \beta) \quad (1)$$

Where ρ is the air density, A is the surface, V_w is the wind speed. Each wind turbine is defined by its own power coefficient which is a nonlinear function depends on the pitch angle β and the tip speed ratio λ .

The power coefficient can be represented as [6]:

$$C_p(\lambda, \beta) = c_1 \left(\frac{c_2}{\lambda_i} - c_3 \beta - c_4 \right) e^{-\frac{c_5}{\lambda_i}} + c_6 \lambda \quad (2)$$

$$\frac{1}{\lambda_i} = \frac{1}{\lambda + 0.08\beta} - \frac{0.035}{\beta^3 + 1}$$

Where $\lambda = \frac{\omega_r R}{V_w}$ and $c_1 = 0.5176$; $c_2 = 116$; $c_3 = 0.4$;
 $c_4 = 5$; $c_5 = 21$; $c_6 = 0.0068$

The model of the dynamic wind turbine drive-train system assumed in several papers [1], [5], [6] is modeled through two mass drive-train system where the low shaft speed is the result of the torsion and friction effects. In order to reduce the model of the system and have a simple structure, the wind turbine is modeled as one mass drive-train system by conceding that the low shaft speed is quite rigid. Fig. 1 shows the reduced single mass drive-train model.

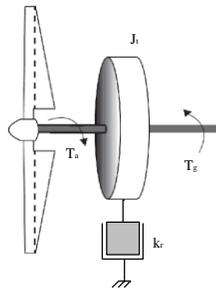


Fig. 1. One Mass Drive-Train Model.

The dynamic characterized by first order equation can be expressed as:

$$\dot{\omega}_r = \frac{T_a}{J_t} - \frac{T_{em}}{J_t} \quad (3)$$

Where J_t is the inertia, ω_r is the rotor speed and T_{em} is the electromagnetic generator torque.

B. Permanent Magnet Synchronous Generator Modeling

The mathematical model of permanent synchronous generator is developed in the *Park d-q* rotation frame linked to the rotor [5].

$$\begin{cases} v_d = R_s i_d + L_d \frac{di_d}{dt} - L_q i_q \omega_e \\ v_q = R_s i_q + L_q \frac{di_q}{dt} + L_d i_d \omega_e + \omega_e \phi \end{cases} \quad (4)$$

Where v_d and v_q are the voltages, i_d and i_q are the currents along the d and q axis respectively, R_s is the stator resistance, $L_d = L_q$ are the inductance of the generator, ϕ is the permanent magnetic flux, $\omega_m = \frac{\omega_e}{p}$ is the electrical rotating speed of the generator in which p is the number of pole pairs. The electromagnetic torque equation can be written as follows

$$T_{em} = \frac{3}{2} p [(L_d - L_q) i_d i_q + i_q \phi] \quad (5)$$

The difference between the d and q mutual inductance tends to zero for a direct drive multiple PMSG [6]. Then the electromagnetic torque depends only on the q axis current. Eq. (5) can be reduced to Eq. (6):

$$T_{em} = \frac{3}{2} p i_q \phi \quad (6)$$

C. Converter Model

The generator side converter (GSC) is a rectifier which is used to control the torque and speed. The three side converter connected to the output of PMSG is presented in Fig. 2.

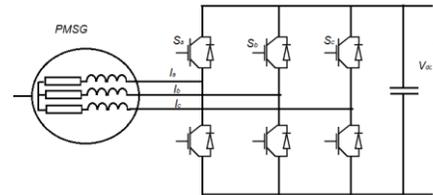


Fig. 2. PMSG with GSC.

According to Fig. 2 the three phase voltage is written as follows [7]:

$$\begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} = R \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + L \frac{d}{dt} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} \quad (7)$$

Where v_a, v_b and v_c are the applied voltages at the machine terminals. They are given by the following equation:

$$\begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} = \frac{1}{3} v_{dc} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} S_a \\ S_b \\ S_c \end{bmatrix} \quad (8)$$

Where S_a, S_b, S_c are the switching variables of the converter, v_{dc} is the DC link voltage.

D. Control Objectives

The objective of the wind turbine control is the tracking of the optimal speed reference that guarantees the optimization of the wind power capture. So the power coefficient must be maintained at its maximum value C_{pmax} which is obtained at an optimal tip-speed ratio λ_{opt} and a specific pitch angle with

$$\lambda_{opt} = \frac{\omega_{opt} R}{V} \quad \text{that } \omega_{opt} \text{ is the optimal rotational speed.}$$

If the conditions optimal are required, the maximum power output is given through [6]:

$$P^{opt} = \frac{1}{2} \rho \pi S C_{pmax} \frac{\omega_{opt}^3 R^3}{\lambda_{opt}^3} \quad (9)$$

However, MPPT control is ensured by maintaining the optimal relation between the generator's speed and the torque without using wind speed measurement. From the optimum power given by Eq. (9) the optimum generator torque can be written as follows:

$$T_{opt} = K_{opt} \omega_{opt}^2 \quad (10)$$

$$K_{opt} = \frac{1}{2} \rho \pi R^5 \frac{C_{pmax}}{\lambda_{opt}^3}$$

Where

The main objective of the PMSG controller is the tracking aeroturbine input which is the electromagnetic reference torque.

III. PMSG CONTROL DESIGN

The mostly often used control approach of PMSG known as field oriented control (FOC) presents several advantages such as accurate speed control and good torque response achieved through the d - q current control loop. This control consists in ensuring that the q axis current measured i_q tracks the q axis current reference i_q^* and that the d axis current measured i_d reaches the d axis current reference i_d^* [7], [8], [9].

In order to produce maximum torque, the d axis stator current is maintained at zero. The developed torque is proportional to the q component of the stator current, so the q axis stator reference current is calculated using the turbine MPPT unit. The overall structure of wind turbine-PMSG control strategy is shown in Fig. 3.

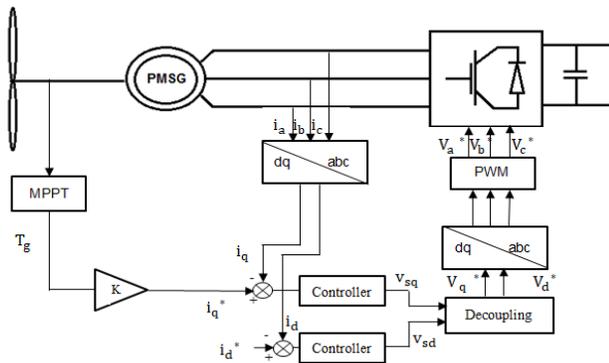


Fig. 3. Wind Turbine-PMSG Control.

The decoupled d and q current loops are expressed by the state equation between the voltage and current on d and q loops and the other components are considered as compensation items. Eq. (4) can be rewritten as:

$$\begin{cases} v_d = v_{sd} - L_q i_q \omega_e \\ v_q = v_{sq} + L_d i_d \omega_e + \omega_e \phi \end{cases} \quad (11)$$

A. Proportional Integral Controller Synthesis

The controller design of this approach is based on the system block diagram as drawn in Fig. 4. The same control

structure loop is applied for the d and q axis current loop control. The transfer function between $i_{d,q}$ and $v_{d,q}$ is given by:

$$\frac{I_{d,q}(s)}{V_{d,q}(s)} = \frac{1}{R_s} \frac{1}{1 + T_e s} \quad (12)$$

Where the stator time constant is $T_e = \frac{L_{d,q}}{R_s}$

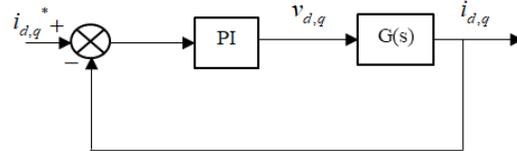


Fig. 4. Current Control Loop.

The Proportional Integral controller is defined as:

$$C(s) = \frac{k_i}{s} \left(1 + \frac{k_p}{k_i} s \right) \quad (13)$$

The parameters of the PI controller are determined through the open-loop pole compensation method as follows:

$$\begin{cases} k_i = \frac{R_s^2}{L_{d,q}} \\ k_p = \frac{L_{d,q} k_i}{R_s} \end{cases} \quad (14)$$

The d and q axis voltages reference v_q^* and v_d^* are generated from the addition of the controllers output to the compensation items, from which the three phase sinusoidal reference voltage is obtained. Thus the control of the stator d and q currents is achieved through the decoupled d and q voltages. PWM is used to generate the switching signal for the power converter. The closed-loop control strategy for generator-side converter is shown as Fig. 3.

B. Internal Model Controller

The vector control of the synchronous generator in Fig. 3 shows that the generated d - q axis voltage is based on the error between the measured and the reference d - q axis currents. The proposed method control of this paper is the use of IMC controller in place of PI controller.

The major advantages of the IMC structure are: the use of IMC controller feedback signal of the difference between the plant model and the reference model, the IMC controller equally ensures the robustness of the system [10].

The design control strategy of the system consists in controlling the d - q stator current with two separate IMC controller loops. The generated d and q current i_{sd} and i_{sq} from the outputs of the IMC controllers are then added to the compensation items in order to compute the d - q reference voltage [14], [15]. The v_d^* and v_q^* are written as:

$$\begin{cases} v_d^* = R_s i_{sd} - L_q i_{sq} \omega_e \\ v_q^* = R_s i_{sq} + L_d i_{sd} \omega_e + \omega_e \phi \end{cases} \quad (15)$$

After the transformation of the measured stator currents from three phase *abc* to *dq* rotating frame, the IMC controllers system operate to minimize the difference between the reference and actual currents on d and q loops.

The IMC scheme block diagram is shown in Fig. 5.

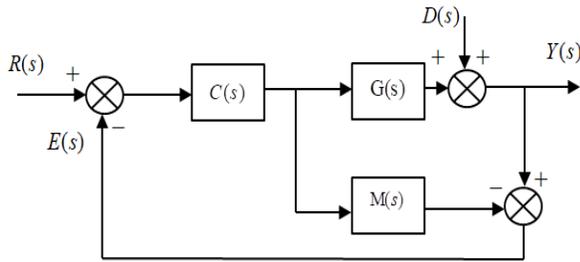


Fig. 5. Standard IMC Structure.

Where $G(s)$ is the mathematical model of the plant, $M(s)$ is the model, $C(s)$ is the model controller, $D(s)$ is the disturbance. $E(s)$ is the information of the disturbance and model plant mismatch as defined as Eq.(16).

$$E(s) = \frac{D(s) + R(s)C(s)[G(s) - M(s)]}{1 + C(s)[G(s) - M(s)]} \quad (16)$$

The IMC structure guarantees the stability of the system for the open loop stable plant. The internal model is perfect, $G(s)=M(s)$ and the closed-loop system is stable if $M(s)$, $G(s)$ and $C(s)$ are stable [11]. Then, in an ideal internal model control is presented as $C(s)=M^{-1}(s)$. However it can be seen that the output of the system cannot reach an input of the system due to a number of reasons [12]:

- a) If is not minimum phase, there are zeros in the right half plane, then is unstable.
- b) There are some parts of the system which are noninvertible.
- c) It is highly sensitive to model errors

In this case and by using the inversion method proposed in [11] which is based on the gain, we achieved an inverse model of the plant model. The IMC controller is given as follows as Fig. 6.

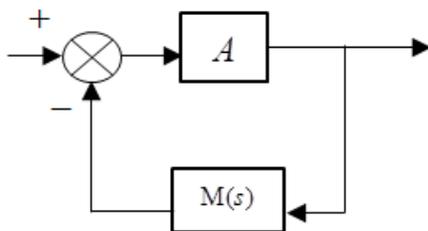


Fig. 6. IMC Controller Structure.

The expression of the internal model controller can be obtained as :

$$C(s) = \frac{A}{1 + AM(s)} \quad (17)$$

Referring to [11],[12],[13] to ensure the stability of the structure proposed, the choice of the gain A must satisfy the condition that the roots of the characteristic equation have negative real parts. The IMC structure can be modified to get a standard feedback control system as shown in Fig. 7. This configuration is more straightforward for implementing the current control loops.

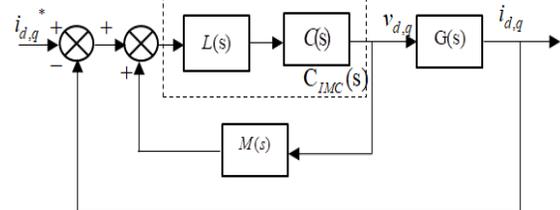


Fig. 7. Modified IMC Structure.

In order to make the system more robust, the controller is raised by a filter. The filter used is the low-pass filter which is given by [13].

$$L(s) = \frac{1}{(1 + \varepsilon s)^n} \quad (18)$$

Where ε is the time constant and n is the order of the system.

The selection of the filter parameter must confirm an acceptable compromise between stability and performance. The adjusting of ε is related to control the stator current of the closed loop response. The filter order should be chosen as appropriately in order to get the fast and robust required system. Then the IMC controller is defined as:

$$C_{IMC}(s) = L(s)C(s) \quad (19)$$

The plant of the IMC control block diagram is $G(s) = \frac{1}{L_{d,q}s + R_s}$ which has strict negative real part root $p_1 = -74.17$, the model is similar to the plant and the filter is taken in the first order.

Then, the expression of IMC controller is defined by:

$$C_{IMC}(s) = \frac{1.42910^{-6}s + 1.0610^{-4}}{5.71610^{-4}s^2 + 5.66910^{-2}s + 1.06} \quad (20)$$

$C_{IMC}(s)$ is stable in open loop because the denominator coefficients of his transfer function are all of the same sign.

IV. VALIDATION RESULTS

The proposed PMSG controller was validated using the SimWindFarm aeroelastic simulator with the parameters of the NREL-5MW variable wind turbine. The NREL-5MW is a

variable speed, variable pitch with a nominal power rating of 5MW, a 126 m diameter, three blades. It is assumed to be coupled to a three-phase PMSG. The main parameters of the wind turbine and PMSG are summarized in Table I.

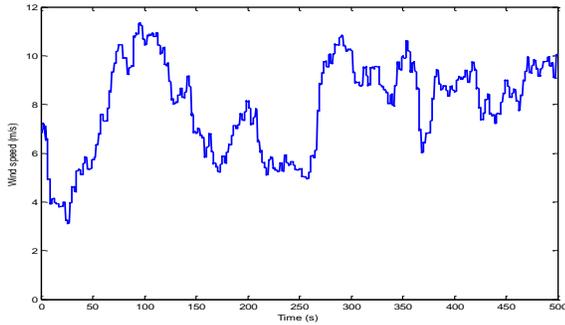


Fig. 8. Wind Speed Profile.

The wind profile used in this study is generated by the simulator. It has a mean value of 7 m.s^{-1} at the hub-height and a turbulence intensity of 25% as shown in Fig. 8.

TABLE I. WT-PMSG CHARACTERISTICS

Parameters	value
Rotor diameter	126 m
Gearbox ratio	97
Hub-height	87.6 m
Maximum power coef.	0.482
Rated speed	1173.7 rpm
Rated power	5 MW
Maximum rotor torque	47,402 N.m
Stator resistance	1.06 Ω
Stator inductance	14.29 mH
Rotor flux linkage	8.6 Wb
Number of pole pairs	5

A. PMSG with IMC Controller Validation

The performance of PMSG control through the designed IMC controller is first investigated. As seen in Fig. 9 the rotor speed is kept around of the optimal reference speed. It is proportioned to the increase of the waveforms of wind speed. The high performance of the controller design is can be seen with the IMC controller which tracks the reference value with reducing steady-state error. Fig. 10 shows the electromagnetic torque which reaches the desired reference value and achieves a good performance. The electrical power resulted is kept near to the aerodynamic power optimal as seen in Fig. 11 with a power coefficient around the constant desired value 0.482, Fig. 12.

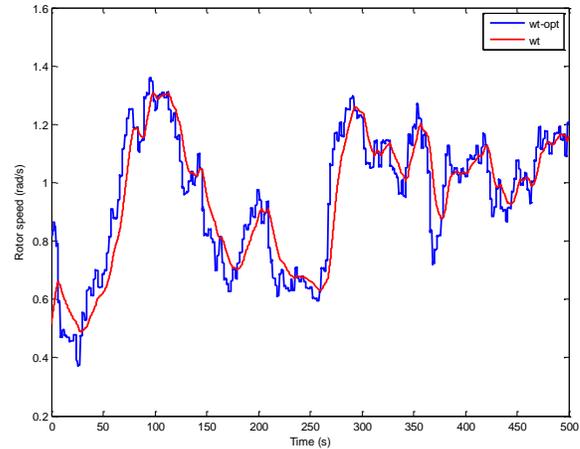


Fig. 9. Rotor Speed Response with IMC Controller.

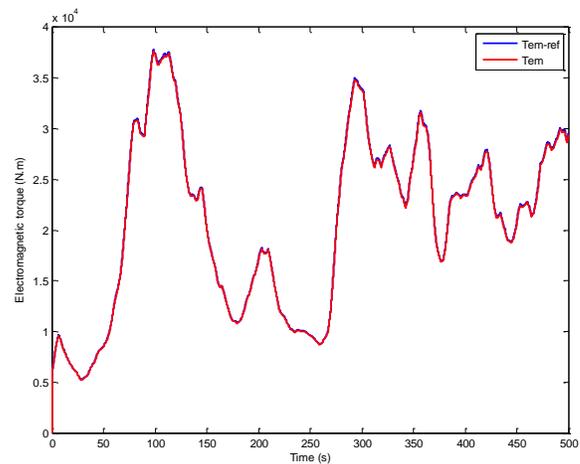


Fig. 10. Electromagnetic Torque with IMC Controller.

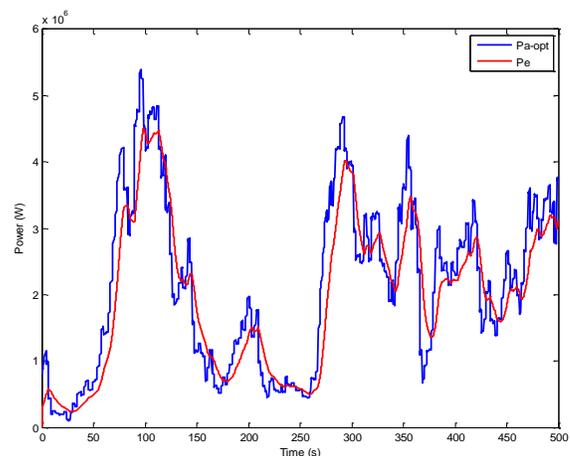


Fig. 11. Power with IMC Controller.

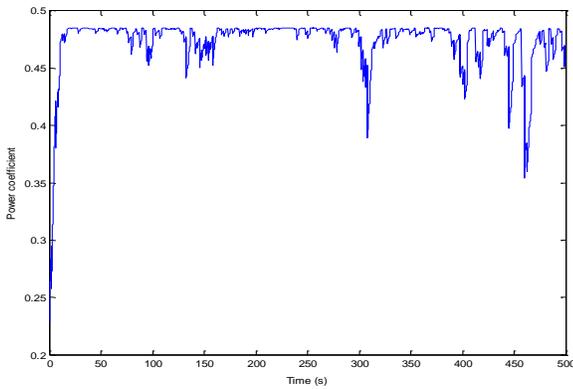


Fig. 12. Power Coefficient.

Fig. 13 shows the dq -axis currents. The direct current component i_d is close to zero, while the quadratic current component i_q is directly related to the reference generator torque. As the desired electromagnetic torque increases so does the q -axis current. It is clearly that the active power is proportional to the quadratic current while the reactive power is only controlled by the direct current. The three-phase abc current as shown in Fig. 14 present a peak of 480A at a rotor speed of 1.3 rad/s. The three-phase voltage developed is presented in Fig. 15 which presents a peak around 6000V at a rotor speed of 1.3 rad/s. The electrical power delivered from the generator converter side shown in Fig. 16 is less than the one calculated through the torque generated. This decrease can be interpreted by the power losses through the converter and mechanical flexible elements.

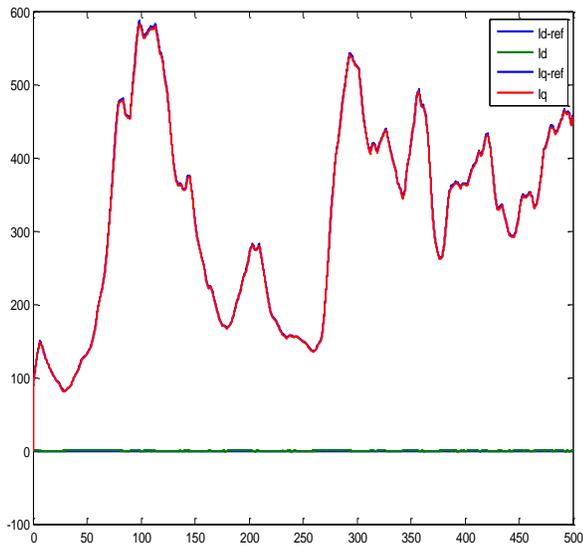


Fig. 13. Direct and Quadratic Current.

In order to test the robustness of the proposed IMC controller, the PMSG parameters R_s and $L_{d,q}$ have been varied with -20% of $L_{d,q}$ nominal value and -50% of R_s nominal value. Fig. 17 and 18 display simulation results for the

parameters variation. It can be seen that the outputs system are able to reach the optimal reference value. So it is clear that the IMC controller has parameters incertitude robustness and it can be seen the effectiveness and the robustness of the designed controller.

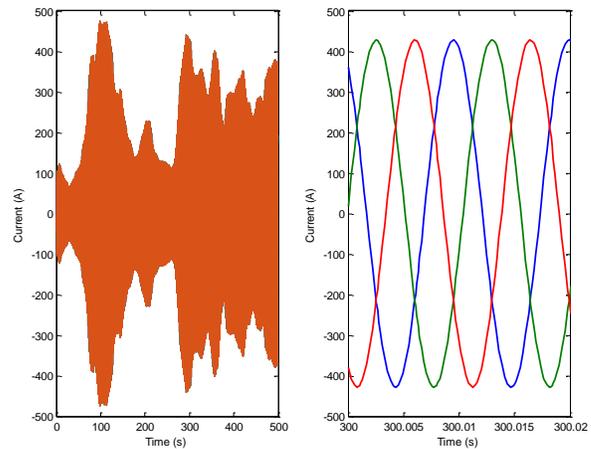


Fig. 14. Three-Phase abc Current.

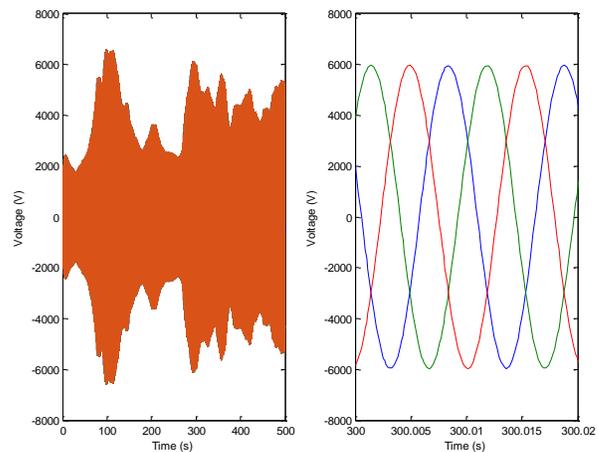


Fig. 15. Three-Phase abc Voltage.

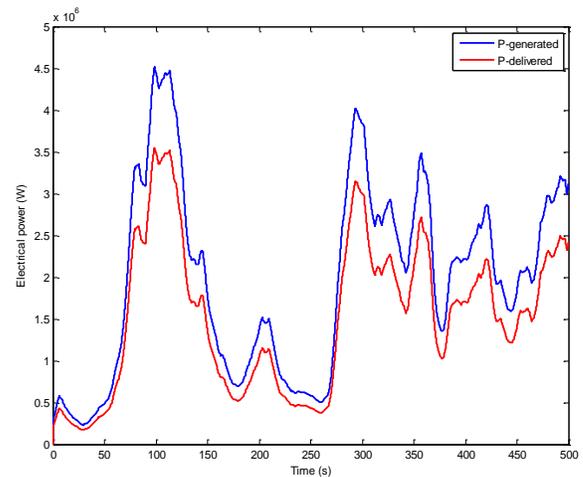


Fig. 16. Electrical Power.

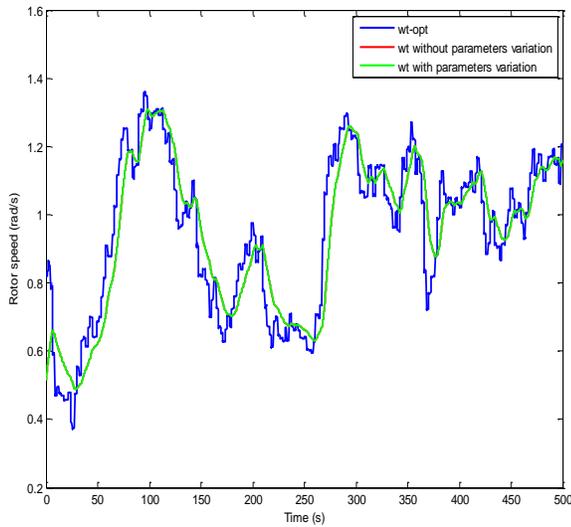


Fig. 17. Rotor Speed

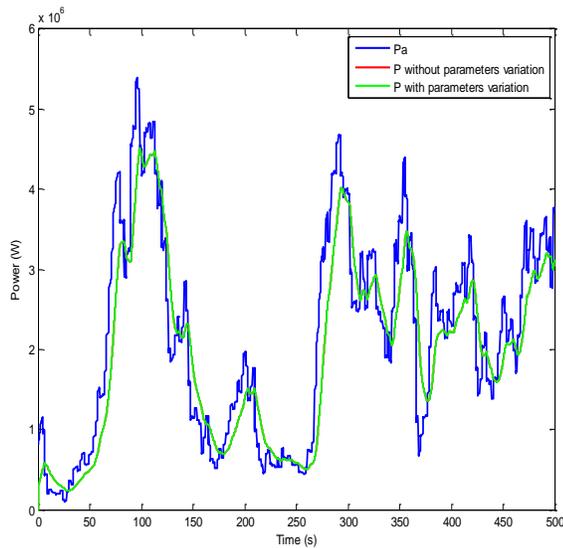


Fig. 18. Power Response.

B. Comparison with PI Controller

The obtained performance of both strategies controllers are shown in Fig. 19, 20 and 21. The selected comparison criteria are the maximum electromagnetic torque, the standard deviation of T_{em} , the maximum power capture and the power efficiency. The simulation results verify that the IMC controller design and the conventional PI controller have the same behaviour. It can be seen from the Table II that the PMSG vector control which made up of two inner-current loops control based on PI action controllers is less effective than the control through the proposed IMC controller. So the IMC method has favourable response robustness and good control effect.

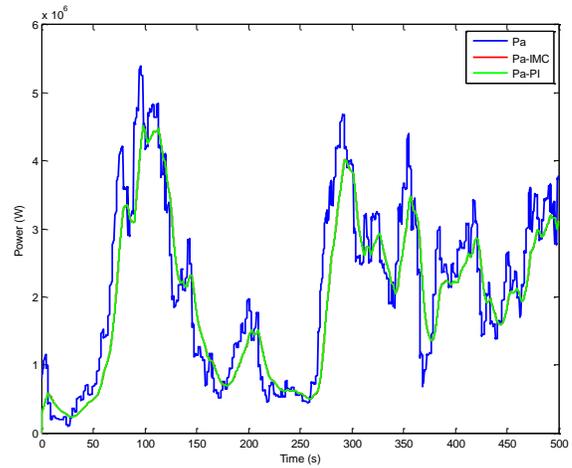


Fig. 19. Comparison of the Power Produced.

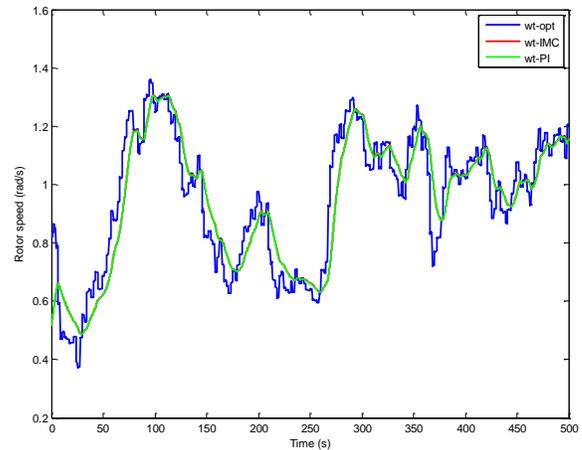


Fig. 20. Comparison of the Rotor Speed Response.

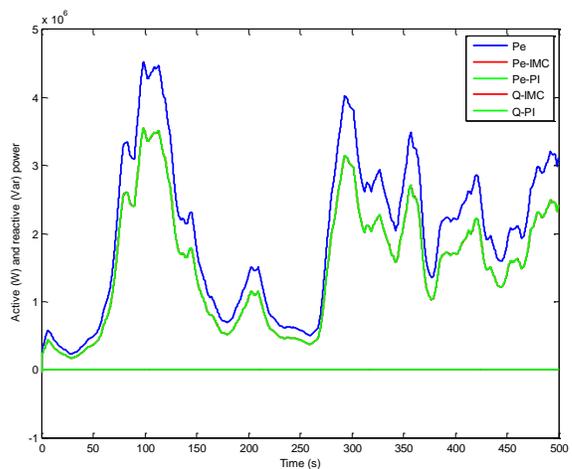


Fig. 21. Comparison of the Active and Reactive Power.

TABLE II. COMPARISON OF THE TWO CONTROLLERS

Controller	PI controller	IMC controller
Maximum power (MW)	4.5150	4.5169
Maximum torque (kN.m)	37.686	37.629
Std (Tem) (kN.m)	8.5897	8.6149
Efficiency %	82.9	83.7

V. CONCLUSION

This paper has dealt with variable speed PMSG wind turbine in order to achieve the objective of maximizing the power energy extract from the wind. Two control strategies are applied to the PMSG: the conventional PI controller and the proposed Internal Model Controller design. The regulating of the generator speed has been provided by the maximum power point tracking. The IMC is applied to the current control to improve performance. The d-q axis currents have been successfully decoupled by the designed control strategy and both of them can follow the reference accurately. The proposed technique is suitable for variable speed wind generation system and ensures the best performance in term of efficiency.

The proposed control design has the advantages of set-point tracking controller and disturbance rejection performances. The simulation results demonstrate the effectiveness and the robustness of the proposed method.

As future research, it needs to take account to the nonlinearity of the power coefficient and integrate a nonlinear controller design for PMSG variable speed wind turbine. Additionally, it is interest to enhance the current study with active and reactive power control exchanged between PMSG and the electrical network during voltage drop.

REFERENCES

[1] Changliang Xia, Yan Yan, Peng Song, and Tingna Shi. "Voltage Disturbance Rejection for Matrix Converter-Based PMSM Drive System Using Internal Model Control" IEEE Transactions On Industrial Electronics, vol. 59, no. 1, January 2012.

[2] Wenchao Meng, Qinmin Yang, and Youxian Sun "Guaranteed Performance Control of DFIG Variable-Speed Wind Turbines" IEEE Transactions on Control Systems Technology Vol.24, pp 2215–2223, 2016.

[3] C.E.Garcia and M.Morari. "Internal Model control-1: a unifying review and some new results." Industrial Engineering Chemistry Process Design and Development, Vol.21, No 2, pp 308-323, 1982.

[4] C.E.Garcia and M.Morari. "Internal Model control-2: design procedure for multivariable systems." Industrial Engineering Chemistry Process Design and Development, Vol.24, No 3, pp 427-484, 1985.

[5] A. Gift IssacÀ , P. K. Senthil Kumara. "Fuzzy Adaptive Internal Model control for the speed regulation of a Permanent Magnet Synchronous motor with an Index matrix converter." International Journal of Current Engineering and Technology, 2014.

[6] Abdullah M.A. , Yatim A.H.M., Tan C.W., Saidur R. "A review of maximum power point tracking algorithms for wind energy systems." Renewable and Sustainable Energy Reviews 16, pp 3220– 3227, 2012.

[7] Guohai Liu, Lingling Chen, Wenxiang Zhao, Yan Jiang, Li Qu. "Internal Model Control of Permanent Magnet Synchronous Motor Using Support Vector Machine Generalized Inverse." IEE Transactions On Industrial Informatics, Vol.9, No.2, May, 2013.

[8] Y.Erramia, M.Ouassaidb, M.Maaroufi "Optimal Power Control Strategy of Maximizing Wind Energy Tracking and different operating conditions for Permanent Magnet Synchronous Generator Wind Farm" Energy Procedia, Vol. 74, August, 2015.

[9] M.Benkahla, R.Taleb, Z.Boudjema "Comparative study of RobustControl Strategies for a DFIG-based Wind Turbine" IJACSA International Journal of Advanced Computer Science and Applications, Vol. 7, No.2, 2016.

[10] Hao Gu, Shihua Li. "Modified Internal Model Control of PMSM Speed-regulation System" The International Federation of Automatic Control Milano, 2011.

[11] A.Dhahri, I.Saidi, D.Soudani "A New Internal Model Control Method for MIMO Over-Actuated Systems". IJACSA International Journal of Advanced Computer Science and Applications, Vol. 7, No.10, 2016.

[12] I.Saidi, A.Dhahri, D.Soudani "IMC Controllers for Uncertain Multivariable Over-Actuated Systems" the International Conference on Advanced Systems and Electric Technologies, pp 346-350, Hammamet-Tunisia, 2017.

[13] A.Hammami, I.Saidi, D.Soudani "Design of an Internal Model Control strategy for Grid side Converter for the Permanent Magnet Synchronous Generator" 5th International Conference on Control and Signal Processing, Tunisia, Vol.26 , pp 145-151, 2017.

[14] D.V.N.Ananth, G.V.Nagesh Kumar, P.V.S.Sobhan, P.Nageswara Rao and N.G.S.Raju. "Improved Internal Model Controller Design to Control Speed and Torque Surges for Wind Turbine Driven Permanent Magnet Synchronous Generator" IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics, 2013.

[15] Zaijun Wu, Xiaobo Dou, Jiawei Chu, Minqiang Hu "Operation and Control of a Direct-Driven PMSG-Based Wind Turbine System with an Auxiliary Parallel Grid-Side Converter."Energies, Vol 6, No.7, pp 3405-3421, 2013.

Impact and Challenges of Requirements Management in Enterprise Resource Planning (ERP) via ERP Thesaurus

Rahat Izhar¹

Research Associate/ ERP Consultant
Department of Software Engineering
Bahria University Islamabad, Pakistan

Saba Izhar³

Teaching Associate
Department of Management Sciences
Bahria University Islamabad, Pakistan

Dr. Shahid Nazir Bhatti²

Associate Professor
Department of Software Engineering
Bahria University Islamabad, Pakistan

Dr. Amr Mohsen Jadi⁴

Associate Professor
Department of CCSE,
University of Hail, KSA

Abstract—Managing requirements efficiently aids the system design team to understand the existence and significance of any individual requirement, there are numerous requirements management practices that benefit in decision making but significantly many lacks to account the important factors that have substantial influence in managing requirements in context of ERP systems in particular. As highlighted comprehensively later in literature review section, requirements management failure is one of the pivotal aspects for the project(s) failure. The prime problem/lacking in software design and development is when it comes to requirements management the most vital thing that gets ignored is thinking before performing activities. As it should be the main step to save time, money and efforts. Further prominence other aspects in this are pivotal value about the software's running in industries, the question arises when their business need ERP system, and when requirements change or new requirements are emerged into the system, what are the obstacles faced and how these obstacles are accomplished. ERP systems are becoming the need of industries nowadays as various industries are facing problems regarding data loss; it is challenging for the owners to fetch all the information when they need it, accounting systems are slower and consuming a lot of time and many other issues likewise. This paper further illustrates in detail the important traits, issues toward businesses may have when ERP is implemented and when requirements are changed or not managed professionally what issues are faced by requirement engineering team and industries and thus how to resolve them.

Keywords—Enterprise Resource Planning (ERP); Product Owner (PO); requirements elicitation; requirements management; change management

I. INTRODUCTION

Requirement is something needed in order to meet an objective. Requirements management demonstrates that organization ensures and meets the needs of its all customers and stakeholders. Requirements Management has a great impact on communication between the team members and all the existing stakeholders as requirements modifications can be

added at any stage of project therefore it's very vital for all the project team members and stakeholders to thoroughly adjust to the requirement changes where it's essential. Firstly business requirements are to be well-defined for ERP, business analysis play a pivotal role in requirements management as it is the responsibility of business analyst to get involved with the stakeholders to create a Requirement Management Plan (RMP) during the planning phase of the project that is ERP system for that specific organization. Requirement management plan includes roles and responsibilities of stakeholder's, requirements management process, requirements prioritization, requirements traceability, requirements change and control, requirement management tools, requirement versioning and many more. The business analyst goes about as an extension amongst business and IT, makes the business requirements that could be comprehended by the development engineers and also disclosing to the clients to the technical know-how of system in detail as required.

Customer relationship ought to be the most important factor while gathering business requirements, as requirements keep on changing throughout the overall requirement engineering process. One of the key factors for a business analyst is to make sure that requirements are properly visible and understood by all the stakeholders. The high level objective of a system under requirements gathering phase is achieved mainly by precision in business requirements. Lucid business requirements are thus obtained in a way that the end results of the project are crystal clear to the stakeholders [9]. A Business Requirements Document (BRD) must be maintained and iteratively checked by the business analyst because it summarizes the answers to the business questions what are the purely business problems that a customer's wants to get solve? For example, what job the customer wants to get accomplish? A business Analyst describes the constraints to a customer on any proposed solution from a business perspective. Thus BRD is formed for the effective communication between internal and external technology providers to provide with the

appropriate solution that satisfies both the customer and business needs [9]. While gathering the business requirements the business analysts deliberates that all the requirements he have gathered are accurate and very relevant to business needs but when these business requirements are thoroughly analyzed these requirements come out to be inconsistent, incomplete and because of these inconsistencies (infected incomplete requirements) in business requirements the overall cost of a project increases [7]. The BRD document ensures the accuracy of business requirements from the business perspective for a customer [8]. When writing the BRD many things get repeated and although analysts take it as a problem but to that he would know that this is what they are expected to achieve in principle. Thus due to clear scope of BRD the clarity between the business objectives and the technical objectives are separated [9]. Business objective defines “what does the organization wants to be or what is the organization mission” whereas “the technical objective provides the ground where the business objectives are met”.

II. REQUIREMENTS MANAGEMENT

Requirements Management describes the features that need to be completed, features include control and track of requirements and changes to requirements at any stage as the project is moved forward. Requirements management has an important portion in contributing better business results. Poor requirements management leads to greater failure chances of overall project [21]. The phases of requirements management process include requirements elicitation, requirements analysis, requirements documentation, requirements validation, change management and traceability as depicted in Fig. 1 below [22].

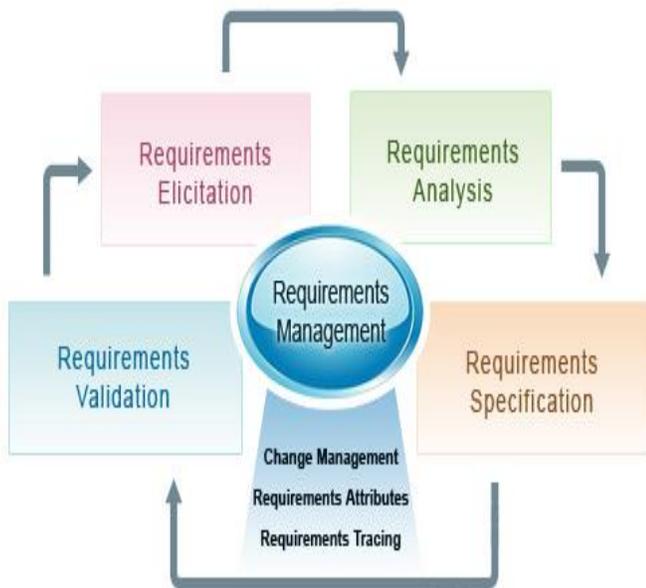


Fig. 1. Requirements Management Cycle [22].

Requirements Management process is iterative throughout the project. As Requirements Management is the continuous process throughout the project lifespan. As requirements are generated at any phase of the project by different stakeholders of the project for example customers could come up with new

or some modifications in requirements, sales could come up with more new added requirements, management with their new or changed requirements, product management may add their more related requirements and so on [3]. There comes communication factor between all the stakeholders when new requirements or change in requirements are being managed. Project managers communicate all the new and existing factors with all the stakeholders. When it comes to communication between all stakeholders it's the responsibility of business analyst to get input from the project manager and create a requirements management plan.

A typical requirements management plan has following components in it [6] [11] [13] [14] [19]:

- *Project overview:* Brief description of project for the readers who haven't seen the project charter. It could be an overview copied from one's development agreement or any other summarized document.
- *The requirements elicitation procedure:* In this segment one will portray the procedure that one will use to evoke, investigate and record the necessities. Thus in this section one infected portraying the prerequisites procedure at a nitty gritty level. That is one might utilize various elicitation systems and there might be various distinctive procedures used [18].
- *Roles and corresponding responsibilities:* This segment records the roles that will be included with dealing the requirements through whatever remains of the task lifecycle. Roles could incorporate the task director, lead expert, customers, and so on. The task supervisor, for example, ought to have the general obligation regarding extension change management of the requirements. Somebody, maybe the lead expert, ought to have general duty regarding the respectability of the requirements all through whatever is left of the lifecycle [13].
- *Tools:* This includes the tools that are going to be used to accomplish the requirements. There are various technologies (tools) that can be utilized to record, oversee and track requirements through the lifecycle. The tools might be as straightforward as MS Word or Exceed expectations, or one may have more advanced requirements discovery, recording and programming tools in this [16] [18].
- *Requirements traceability:* The overall process of tracking/tracing requirements throughout the project lifecycle is described in this section. This procedure must then be supplemented to the agenda to guarantee the best possible following of requirements happens all through whatever is left of the project [21].
- *Change Control:* The formal process that is performed to manage the change in requirements is written in this section. It is expected that the formal scope change process is being implemented throughout the project. If this process is being used then it's applicable to the variations in requirements and if there's no official

modification procedure is being used then a specific formal change process is documented in this section.

- *Approval:* The project manager's approves overall writings of the project and ensures the every document regarding requirements in this section.

Requirements management plan enables the whole project to be on the same page in regard to requirements elicitation and relevant requirements related issues. This could ensure an effective communication between the stakeholders and increasing number of odds of the projects success which comprises customer needs and achieving successful time-to-market [3]. These advantages are particularly valid for substantial activities that include different, internationally disseminated utilitarian units.

III. ERP SYSTEMS

Enterprise Resource Planning (ERP) system helps the enterprise organizations to automate and integrate all their human resource management, financial management, material management, sales and distributions, production planning & control, plant maintenance, document management and workflow, security and system administration activities [5], ERP makes business more profitable. The automation and integration of an organizations core business provides level of control [5]. That is business becomes smarter and strategic decisions are easy to make. It is like umbrella term for complicated type of software setup. It can be customized to entailer one's needs. Business needs specific software solution, as Fig. 2 shows some of the modules/ departments an ERP covers [4].



Fig. 2. Different Modules of ERP System [4] [11].

IV. ERP TECHNICAL DETAILS/ METHODOLOGY

In ERP systems one have to have access to Task Management, Fleet Management, Human Resources, Notices, Expense Management, Purchase Management, Project Management, Invoicing, Stock Management, Dashboards, Detail Reports, 24/7 support and more other business modules. For example one stakeholder has a store this store bonds all his products. When product will be directed ERP will be included in every point of sequence. It will systematically schedule distribution and arrange to ship and deliver, thus observing in transit through convoy supervision and even replenish that product all inevitably. If there's something missed there is not a problem, it can be added or removed because ERP has a feature of customization [2]. ERP is advanced it can be controlled over the mobile phone at very affordable cost. Top five (5) signs one's business needs an ERP [3].

- 1) Different software's for different processes.
- 2) Difficult access to fetch information about one's business.
- 3) System gets slower i.e. accounting takes longer than it takes.
- 4) Sales and customer experiences are suffering.
- 5) The IT (automation) is too complex and time consuming.

Multiple components into one system: ERP optimizes the business, centralizes data, automatic handling, cost reduction, audit trails, business continuity and stability [3]. With ERP single stream of data is possible. Integrates and automates the data management and company's business processes.

ERP promises Accurate data, faster decisions making, lower operations costs. ERP is the tool of managing information which means, A company has its CRM (Customer Relationship Management) something manages all company's orders or warehouse. An accounting system and something filling the gap's in one's information pipeline which could be spreadsheet and manual procedures but none of these negotiation to each other, they do not part any information which impacts proficiency [8]. Efficiency is associated to time and money. At this stage ERP helps one the most as it centralizes all the information in one's organization, by enabling this one can stream line the course of data. As most of our business processes become connected End to End. All additional things become informal that is, single opinion of one's client so one can deliver better provision. Thus our team effort more proficiently as they have the evidences they needed and when one need to analyze that information since it's all in same place, one can report on it any way one want to do so [16].

V. METHODS AND MATERIALS

A. *Critical Achievement Aspects Enterprise Resource Planning (ERP) System Implementation: An Investigative Study in Oman*

[1], this research paper addresses to Critical Success Factors that are raised in Enterprise Resource Planning (ERP) System Application. These are main 10 basic CSF's can help organizations to accomplish fruitful usage of ERP framework. These CSF's are Clear Goals and Objectives, Strategic IT Planning, Monitoring and Evaluation of Performance, User Training and Education, Top Management Support, Vendor Support, Teamwork and Composition, User Involvement, Project Champion and Education on new Business Processes. The authors focuses that focusing in mind the primary objective is to encourage the tenacious achievement of ERP framework and to assurance a greater effect on the business performance; CSFs must be required with the ERP context amid the practice of ERP outline to commercial requirements. From this work, it can be easily determined that ERP framework drives basically three primary execution phases of framework life cycle and that includes pre-usage arrange, implementation stage, and post execution organize.

B. *A Commercial Procedure Modeling – Enable requirements Engineering Framework for ERP Implementation*

In [2], the authors describe that productive implementation of an ERP framework strongly relies upon the precise meaning of primary functional requirements. The following paper also discusses a case study that portrays and dissects the advantages of the use of a prerequisites designing system to help Enterprise Resource Planning (ERP) advancement. This system further consolidates the innovation driven and the procedure driven methodologies for prerequisites investigation and execution. Particular business process displaying strategies improve the system and help the detailing of the utilitarian determinations of the ERP framework and the administration of prerequisites.

The connected structure associates necessities designing with the ERP improvement life cycle and utilizes a business procedure demonstrating approach keeping in mind the end goal to help both the RFI and RFP creation. The approach shelters the hierarchical view, the procedural view and the data innovation see all together to characterize data innovation resolves related with the operation of the organization and the requirements of its partners. The proposed arrangement of exercises in manual, IT-empowered and robotized grants the simple disconnection of those that prompt useful determinations and empowers the age of more explanatory definitions in light of the accessible procedure streams.

The procedure displaying approach ensures that necessities definition isn't a self-assertive organization of various partners' conclusions however a cognizant and facilitated choice serving the key targets of the enterprise and its vision of its future procedures. Also, it bolsters it to characterize the level of selection of its procedures to ERP usefulness before the RFP procedure, guaranteeing that it will work as per its interesting business forms after the ERP usage. The making of the structure with the above attributes fulfils the principal inquire about goal.

C. The Design of Manufacture Components of ERP System based on Requirements Engineering of Electrical Manufacturing Services

[3], this research paper describes that to lead the elicitation procedure successfully, the organization's partner ought to have a similar comprehension and information on the center term. In the examination and transaction of useful prerequisites, there has been a concentrated exchange between the designer and client on clarifying the hole between as-is and the to-be show. This progression is additionally a tedious advance after the elicitation step. This progression will be quicker if the organization has utilized the data framework in supporting the business procedure. With the involvement in utilizing the data framework application, the clients will see progressively the rationale of the framework in supporting the business procedure.

The author specifically addresses the Production Department in PT.TDK (one of the Electronic Manufacturing Service Companies located in Serang City Indonesia) has never utilized the data framework application previously; to date the business procedure is just upheld by spreadsheet. A business work in the generation territory that we have broken down is the gathering procedure. PT. TDK is an Electronic Manufacturing Service organization whose center business is to process get together of merchandise so there is no generation procedure from crude materials. Best practice as capacities required in outlining ERP frameworks in this investigation construct in light of business process on the PT. TDK. The capacities are coordinated isn't just ERP best practice; however there are particular capacities identified with government controls. Best practices in this investigation can be connected to alternate EMS organizations by considering the particular capacities.

D. Modules and Investigation Scheme of Enterprise Resource Planning Requirements in Small and Medium Enterprises

In research work [4], authors describes that an ERP requirements investigation scheme for ERP system development in directive to yield the appropriate ERP system utilities for small and medium enterprises by cutting down the parts of necessities and the association of the commercial procedure demonstrating, a few fundamental ideas are given and the technique for the procedures analysis and modelling is additionally communicated. The method used for the ERP requirements matching the SME's requirements is systematic. By providing approaches a) the post perceptible link within functionality and ERP requirements b) an international go through of ERP system working.

E. Formalizing Requirements in ERP Software Implementations

In [5], authors address to the research problem of functional and data challenges in ERP. They point out two systems by applying bipartite graphs and determine the matching functionalities through analysis. In spite of the fact that execution scope develops exponentially, putting the greater part of the spaces required into one single bi-chart definition was most certainly not conceivable; authors apply a gathering of sub-charts.

A bipartite diagram G can be considered as an item $G \sim H \times K2$ of a chart H with the relating finish chart K2 like considering a significantly number g into an item $g = h \times 2$, with the exception of that for charts the factorization require not be one of a kind. This exchange off was convenient in running iterative improvement forms. Finally they describe in a case study the event bidding process, analysis of legacy systems and agile product line implementation and provide the results into one signal bi-chart.

F. Enterprise Resource Planning Classifications: Requirements Analysis, Assessment and System Choice – ERP Decision Analysis Using a Game Theory Approach

In [6], authors propose some of those amusements and explore the effect of specific practices in the ERP prerequisites examination process, ERP assessment process and with ERP framework decision, situated with regards to two genuine cases. They describe three organizations that have done prerequisites examination, assessed and picked between various ERP frameworks. Those cases are utilized to delineate different potential recreations that can be played. Also abridges a few prerequisites investigation programming and qualities of that product, for example, a solitary component can "execute the arrangement." Plus describe ERP requirements examination and coming about amusement playing practices, for example, influencing utilization of "bargain slaughtering" requirements or modifying their needs on requirements with a specific end goal to impact to framework decision. As they examine how firms assess diverse ERP frameworks and a portion of the amusement playing practices that can come about, for example, changing branch assessment weights to impact the framework decision. In this way, few issues can happen in framework decision through plan setting in various branches, altogether impact framework decision. And ends up by Stacking the Deck" that is picking individuals from the assessment group to finds the solution that was to be investigated.

G. ERP Requirements for Supporting Management Decisions and Business Intelligence

In research work [7], the author presents a hypothesis test for finding the essentialness of directors needs in the field of choice help and Business Intelligence, 11 things were separated as noteworthy needs. The discoveries of the exploration are the contrast between the measures of utilizing mechanized data frameworks during the time spent basic leadership among various levels of administrators. The discoveries demonstrated the profound contrast in the necessities of choice bolster among various levels of administrators. The analysis describes that if tools for reacting to choice help and BI needs are actualized in ERP frameworks, they ought to be tweaked deliberately as to the level of administration that utilizations them and their profundity and points of interest ought to be set by the methodologies of various administrators towards computerized frameworks. With a specific end goal to react to the reported needs, different devices and BI arrangements were likewise prescribed.

To organize the usage of these arrangements in ERP frameworks, the weights of criteria (requirements and necessities) and need for executing the choices (BI arrangements) were computed from an official viewpoint in

light of requirements as criteria of basic leadership and in light of anthropic skills and intuitive improvement strategy .The outcome can be a suitable guide for providers of these frameworks and associations that expect to actualize them.

H. Improving ERP Requirements Specification Process of SME's with a Customer- Centered Analysis Method

In [8], the authors initially highlight the qualities of Small Medium Enterprises (SMEs') ERP ventures and through contextual investigations examine how they could be developed all the more viably. They describe that propose propelled strategy for SMEs' ERP prerequisite detail that incorporates operational, relevant and hazard examination. Together these investigations give an entire top to bottom depiction of organization business process advancement, determine the prerequisites for the new ERP framework, and recognize the confinements and dangers identified with the ERP venture.

The consequences of a contextual analysis in which the strategy has been connected, bolster advance improvement towards an all-encompassing and multidisciplinary approach in the ERP necessities particular procedure of SMEs. Further they proposed a new vendor and software independent method for ERP requirement specification of Small Medium Enterprises. The proposed technique has clear focal points since organization particular prerequisites are considered and is applied in different SME's to discover its appropriateness for the organization and its business.

VI. REQUIREMENTS MANAGEMENT IN ERP

A comprehensive and systematic analysis of the research work accomplished in context of 'Requirements Management in ERP' by the different researchers, practitioner is highlighted in detail in Table I below.

TABLE I. COMPREHENSIVE ANALYSIS OF CONTRIBUTIONS AND LIMITATIONS OF REQUIREMENTS MANAGEMENT WITH RESPECT TO ERP VIA LITERATURE

Author	Year	Overall Contribution	Limitations
Nikolaos A. Panayiotou, Sotiris P. Gayialis, Nikolaos P. Evangelopoulos [2]	2014	Presentation of ERP reference models gathered with organization processes can lead to successful implementation of ERP system.	Single case study provided in a representative manufacturing company, but further research in other industries can be conducted to find out other requirements and success factors for implementing ERP system.
Kursehi Falgenti, Chandra Mai, Said Mirza Pahlevi [3]	2015	Presentation of ERP reference models gathered with organization processes can lead to successful implementation of ERP system.	Single case study provided in a representative manufacturing company, but further research in other industries can be conducted to find out other requirements and success factors for implementing ERP system.
Yousef Khaleel, Anmar Abuhamdah, Mutaz Abu Sara, Bassam Al-Tamimi [21]	2015	Proposed conceptual components of ERP requirements required for initiating ERP system functions.	ERP system required only for small medium enterprises.
Talukdar S. Asgar and Tariq M. King [5]	2016	Proposed an approach that uses bipartite graphs to formalize ERP requirements using a case study author also describes how this model can be used to formalize large scale projects that are an ERP implementation.	Beneficial for many ERP software engineering process. All domains in a Single graph was impossible so multiple bi-graphs that is sub graphs are used. Future work involved development of a tool that may be able to define multiple domains.
Daniel E. O'Leary [7]	2000	Reviewed some games and compared with the impact of certain ERP requirement analysis and requirement evaluation process in ERP systems.	Focus was only on two branches that is two requirements at a time. But the analyses can be extended to more requirements at a time.
Mehdi Ghazanfari1, Mostafa Jafari and M. T Taghavifard [8]	2009	Survey of fundamental expectations of managers for various levels of ERP systems and provided the significance of requirements for business intelligence for ERP systems proposed using statistical analysis tool.	Focused on the tools for decision support and BI needs in an ERP system that must be customized precisely
Inka Vilpola1; Ilkka Kouri [9]	2005	Provided the characteristics and case study how SME's ERP projects can be evolved more successfully. Methods for SME's ERP requirement specification that has operational, contextual and risk analysis.	Specifically focused on characteristics and provided methods for only small medium enterprise ERP systems.

Ali Tarhini ¹ , Hussain Ammar ¹ , Takwa Tarhini & Ra'ed Masa'deh [10]	2015	Provided a systematic review of 35 research articles published on critical success factors of ERP system implementation between 2000 and 2013. 51 CSF's in ERP implementation are suggested important.	Reviewed research papers only and provided several CSF's for more efficient and successful ERP systems.
Jan Mittner, Alena Buchalcevova [11]	2014	Surveyed that small companies are not satisfied with software systems. Based on literature analysis, market and own experience first version of requirement specification of ERP system for small companies was created and validated.	Focused specifically on small companies' requirements.
Jun Li a, Tangtang Xie b and Shuang Du [12]	2011	Proposed five functions that an ERP system must have for small and medium publishers for increasing the flexibility and efficiently solving many problems like business problems.	Limited to only small and medium publishers to increase the efficiency and flexibility of ERP system modules.
Ahmad Saleh Shatat [10] and earlier by Rolland and Prakash [20]	2015	Important role of critical success factors CSF's in ERP system.	Limited to only Critical success factors in ERP and specially study was investigated in only Oman.

VII. ERP BUSINESS FACTORS

Early Researches [3] [9], [10] published research work on ERP business factors may include the following critical success factors (CSF's):

- Top Management support and inclusion
- Project Competence and association
- Clear objectives destinations and extension
- User preparing and training
- Change Management
- Business process Reengineering
- Effective communication
- User association
- Data examination and transformation
- Consultant
- Architecture choice
- Minimal customization
- Project Management

Implementing an ERP system in an organization is a very crucial process. Requirement engineering team must know the overall details and relevant factors of the system intensively. Various components must be considered and requirements must be gathered before implementation of the ERP system [8]. Requirement engineering team must go through all the departments in an organization and then decide which component must be implemented. Critical success factors must be considered perfectly so that a successful ERP could be implemented.

Different organizations have different requirements and according to the thorough analysis of those requirements ERP system takes place [8]. Fig. 3 shows some common modules an ERP have for medium and small organizations.

When an ERP is implemented it must fit well with the company's present operations and structures and it ought to in like manner have the limit to convey on key execution and gainfulness objectives that the company's current frameworks cannot convey [10]. This is a troublesome demand as it is furthermore the inspiration driving why ERP has a higher failure rate. It can be a calling completing endeavor for a few of the directors if an ERP foundation doesn't go well. Thus it is the responsibility of requirements engineering team to get involved with the organization so that a successful ERP can be implemented in place.



Fig. 3. Components of ERP [4] [10].

VIII. EFFECT OF CHANGE MANAGEMENT TO ERP

When an ERP system is getting in to place in an organization it is very difficult for the staff to grasp the whole system easily at the start. Many of the staff members do not agree with changing the environment because they were comfortable with the old system and start feeling about the new system as a burden. It is the responsibility of the high authority (management) to provide trainings to the staff of new system and explaining the reason why the old system is discarded and how the new system will benefit them. The organizations that are already having ERP system in place and are not satisfied with that ERP system it is difficult to implement in such

organizations because it takes a huge amount of time to analyze their system and implementing new system accordingly. Whereas those organizations that are not having any ERP system and doing work on excel sheets and using different software's for different processes it is much easier to place an ERP system in such organizations because modules can be directly implemented and get integrated with other departments. Fig. 4 illustrate a snapshot of requirement elicitation regarding an ERP system that is how requirements elicitation is used in ERP system's to confirm the requirements of stake holders and requirement engineers.

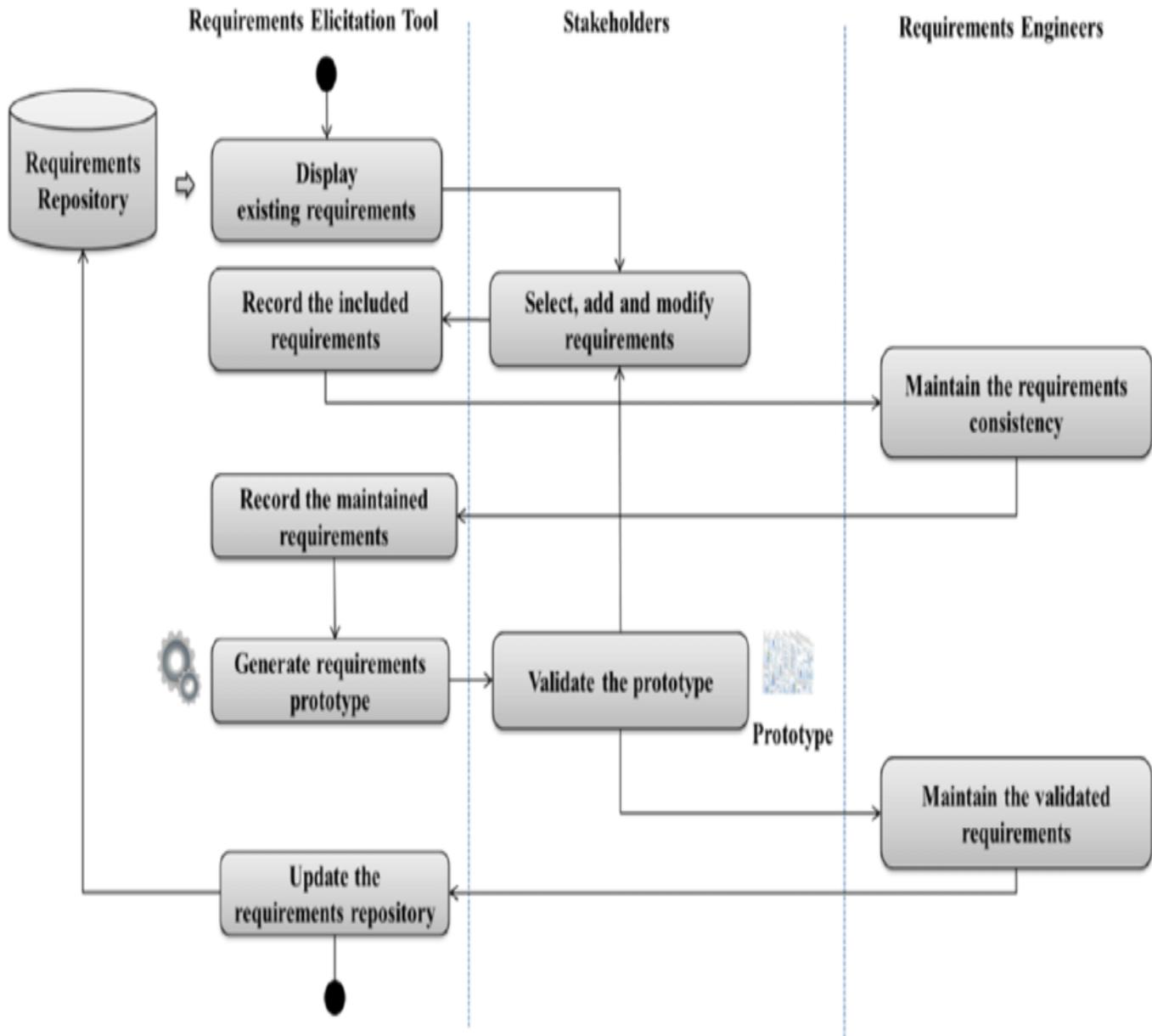


Fig. 4. Requirements Elicitation in ERP System [20].

IX. ERP CHECKLIST

The prime idea is that whenever an enterprise organization requires a system that is integrated and fulfilling business needs properly, requirement engineering teams plays a vital role while a new system is being built for an organization. For example ERP covers all the departments in an enterprise organization that is it cover the entire financial accounting module, human resource management module, manufacturing and distribution module, material management module, sales and inventory module, supply chain module, customer Relationship module, integration capabilities, and support. Further these modules are elaborated and checklist for each module is shown in figures below [15] [17] [19].

A. Financial Accounting Module

Financial module must cover's all the functionalities that are related to financial activities. The task of requirement engineer's here should be to make a proper checklist for the stakeholders to let them know about what each module essential factors may have to be there. Fast analysis and proper accuracy makes fewer amount of errors and allows the organizations to have proper knowledge of its financial health. The vital thing to consider on the grounds that ERP computerizes a few departments in finance accounting, for example, finances, invoicing and planning, that more often than not take many staff hours on a month to month premise [4]. Fig. 5 shows a requirements checklist that must be checked by requirement engineer for a successful ERP financial accounting module.

- Payroll
- Accounts Receivable
- Accounts Payable
- Invoices
- Deposits
- Budgeting
- Bank Reconciliation
- ERP to Bank Interfaces and Reconciliation
- Cash Flow Management
- Automatic Cost Calculation
- Cost Calculation Mode Selection
- Cost Analysis
- Advanced Allocations
- Budgeting
- Expense Management
- General Ledger
- Invoicing/Billing
- Multiple Currencies Support
- Multiple Languages Support
- Regulatory Compliance Support

Fig. 5. Financial Accounting Checklist.

B. Human Resource Management Module

The automation of HRM module reduces the human errors like tax calculations and benefits the administration by

monitoring all the individual productivity and managing hiring data. This will profit the organization increasing its efficiency and financial losses are reduced that were due to imprecision [7]. Fig. 6 shows a requirements checklist that must be checked by requirement engineer for a successful ERP HRM module and its relevant features.

- Benefits administration
- Enterprise Compensation Management
- Human Resource Management
- Payroll Integration
- Employee Performance Management
- Tax Administration:
- Personal Income Tax
- Employee Self Service
- Manager Self Service
- Talent Management
- Time and Labor/Attendance
- Time Card Management

Fig. 6. HRM Checklist.

C. Manufacturing and Distribution Module

The signs of a successful ERP are that it must be able to manage manufacturing and distribution processes to spare time and cash by amplifying staff hours [15]. Manufacturing and distribution process is useful because it covers entire tasks of distribution scheduling, production control and quality analysis and saves the extra hours that the staff may consume without this module. Fig. 7 shows the checklist that must followed by requirement engineer for successful ERP manufacturing and distribution module.

- Capacity Requirements Planning
- Production Cost Analysis
- Production Control
- Process Synchronization
- Product Data Management
- Engineering Change Management
- Forecasting
- Job Costing
- Master Production Scheduling
- Quality Assurance Management
- Quality Monitoring
- Quality Analysis
- Resource Planning and Management

Fig. 7. Manufacturing and Distribution Checklist.

D. Material Management Module

All the business activities related to stock for example deliveries, monitoring and orders are covered in this module. Business Intelligence tools in ERP helps the organizations to predict sales trends to control stock level. The functionalities

and features of this module help the retailers and manufacturers to maximize storage space, get rid of unproductive inventory and meet revenue goals [12]. Fig. 8 shows a requirements checklist that must be checked by requirement engineer for a successful ERP Material management module.

- MRO
- Availability of Materials
- Availability of Stock
- Optimum Store Inventory
- Planning

Fig. 8. Material Management Checklist.

E. Sales and Inventory Module

This module covers all the sales and order management that is investigations sales information, computerizes exchanges, screens execution, tracks costs and even oversees arranges readiness and client credit. This will allow the real time location of any item available in the inventory. The Automatic price/profit calculation will help to save the time and effort and increase efficiency and accuracy of the business. Fig. 9 shows a requirements checklist that must be checked by requirement engineer for a successful ERP Sales and Inventory module.

- Customer Credit Management
- Pricing
- Picking
- Quotations
- Sales Order Processing
- Capable to Promise Inventory
- Bill of Materials
- BOM Reporting
- Credit Card Processing
- Online Transaction Reporting
- Sales Order Management

Fig. 9. Sales and Inventory Checklist.

F. Supply Chain Module

Managing various processes like procurement processes, material resource management, supply chain management, complex processes, distribution and delivery planning and reduces the risk of human error. This module of ERP is mainly beneficial for distributors, retailers and manufacturers. Fig. 10 shows a requirements checklist that must be checked by requirement engineer for a successful ERP supply chain module.

- Logistics Management
- Logistics Planning
- Procurement
- Sourcing
- Packaged Integration to Back Office Applications
- Demand Planning
- Distribution Management
- Event Management
- Manufacturing Execution System
- Management of Resources
- Scheduling
- Dispatch of Orders
- Execution of Orders
- Collection of Production Data
- Production Performance Analysis

Fig. 10. Supply Chain Checklist.

G. Customer Relationship Module

As the customer satisfaction is the most important part of any business success. ERP must provide the Comprehensive customer tools in Organizations for accomplishment of customer satisfaction and robust sales. Fig. 11 shows a requirements checklist that must be checked by requirement engineer for a successful ERP customer relationship module.

- Customer Account Management
- Customer Account Dashboard
- Contact Management
- Automated Marketing
- Marketing Resource Management
- Pricing and Tracking
- Cost Summaries
- Service Management
- B2C Commerce
- B2B Commerce
- Order Management
- Web Storefront Tool
- Email Tools
- Email Integration
- Case Management
- Customer Self Service
- Marketing Campaign
- Rental Management
- Sales Management
- Mobile Access
- Sales Force Automation

Fig. 11. Customer Relationship Checklist.

H. Integration Capabilities

ERP integration ensures that the employees of the organization use the system quickly and effectively. Once ERP system is implemented it allows analyzing data, generating real time reporting, performing business tasks and integration of any new application is easy. Fig. 12 shows a checklist of integration capabilities of an ERP system.

- ✓ Module Integration
- ✓ Hardware Integration
- ✓ Application and Software Integration
- ✓ Integration with Handheld Mobile Devices
- ✓ Integration with portable scanners
- ✓ Integration with fax server
- ✓ Integration with barcode generator

Fig. 12. Integration Capabilities Checklist.

I. Support

The most important thing for an ERP provider is that it ensures and supports the user support. As ERP system is a huge system it requires training, troubleshooting and repairs when implemented. Fig. 13 shows the support and ERP must provide after implementation.

- ✓ Training
- ✓ Maintenance
- ✓ Phone
- ✓ Email
- ✓ Chat/IM
- ✓ Helpdesk Support
- ✓ Forum/Community Support

Fig. 13. Support Checklist.

These all modules corresponding checklists may help the ERP system requirements engineers and developers to make an effective system that is having all the functionalities and features included in each module and developing a successful ERP for an enterprise organization to meet the customers as well as business needs.

X. CONCLUSION

Requirements emerge throughout the software design and development processes and thus requirements are needed to manage with utmost significance, especially when the scenario is that of ERP Systems. As disused and emphasized in this work, there are various requirements management practices, methodologies proposed and been trailed but most of them fail to justify those classical factors (metrics) [21] that influence the overall design and implementation of software product (attributes) being developed. The pivotal idea thus here is that the system requirements must be gathered completely by requirements engineering team before designing and developing a system. ERP system(s) are becoming prominent/ in need in this era for enterprise organizations. An organization must be thoroughly analyzed that is the requirements engineering team must create all the related documents and complete work to develop a successful system for the concerned organization. All the functionalities and features (as

highlighted in section IV of this paper) of the modules discussed above must be gone properly to ensure that the ERP system is complete and consistent for that particular organization.

REFERENCES

- [1] Ramayah, T., Arokiasamy, S., & Eri, Y. (2005). Critical Success Factors in Enterprise Resource Planning (ERP) system Implementation: Results from an Exploratory Study. 2nd International Conference on Business & Economics, "Capitalising the Potential of the Asian Integrated Market, 28(1), 2005.
- [2] Panayiotou, N. A., Gayialis, S. P., Evangelopoulos, N. P., & Katimertzoglou, P. K. (2015). A business process modeling-enabled requirements engineering framework for ERP implementation. *Business Process Management Journal*, 21(3), 628–664. <http://doi.org/10.1108/BPMJ-06-2014-0051>
- [3] Falgenti, K., Mai, C., & Pahlevi, S. M. (2016). The design of production modules of ERP systems based on requirements engineering for Electronic Manufacturing Services company. 2015 International Conference on Information Technology Systems and Innovation, ICITSI 2015 - Proceedings, (March 2016). <http://doi.org/10.1109/ICITSI.2015.7437709>
- [4] Khaleel, Y., Abuhamdah, A., Sara, M. A., & Al-Tamimi, B. (2016). Components and analysis method of enterprise resource planning requirements in small and medium enterprises. *International Journal of Electrical and Computer Engineering*, 6(2), 682–689. <http://doi.org/10.11591/ijece.v6i2.8372>
- [5] Asgar, T. S., & King, T. M. (2016). Formalizing Requirements in ERP Software Implementations. *Lecture Notes on Software Engineering*, 4(1), 1–40. <http://doi.org/10.7763/LNSE.2016.V4.220>
- [6] Sehrish Alam, Dr. Shahid Nazir Bhatti, Impact and Challenges of Requirement Engineering in Agile Methodologies: A Systematic Review, *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(4), 2017. <http://dx.doi.org/10.14569/IJACSA.2017.080455>
- [7] O'Leary, D. (2000). Game Playing Behavior in Requirements Analysis , Evaluation , and System Choice for Enterprise Resource Planning Systems. *Icis*, (May), 385–395 Paper 35.
- [8] Ghazanfari, M., Rouhani, S., & Jafari, M. (2009). ERP Requirements for Supporting Management Decisions and Business Intelligence. *The IUP Journal of Information Technology*, V(3), 65–84.
- [9] Vilpola, I., & Kouri, I. (2005). Improving ERP Requirement Specification Process of SMEs with a Customer-Centered Analysis Method. *Frontiers of E-Business Research 2005*, (October), 140–151.
- [10] Tarhini, A., Ammar, H., Tarhini, T., & Masa'deh, R. (2015). Analysis of the Critical Success Factors for Enterprise Resource Planning Implementation from Stakeholders' Perspective: A Systematic Review. *International Business Research*, 8(4), 25–40. <https://doi.org/10.5539/ibr.v8n4p25>
- [11] Mittner, J., & Buchalceva, A. (2014). The ERP System for an Effective Management of a Small Software Company – Requirements Analysis. *Journal Of Systems Integration*, 5(1), 76–87. <https://doi.org/10.20470/jsi.v5i1.187>
- [12] Li, J., Xie, T., & Du, S. (2011). Requirements analysis on flexibility of ERP system of medium and small publishers. *Procedia Engineering*, 15, 5493–5497. <https://doi.org/10.1016/j.proeng.2011.08.1019>
- [13] Rashid, A., Zakeriya-Nas, Shami, M. U. D., Muhmood, U., Naila-Gul, & Ceylan-Oklu. (2009). Windmills & CSFs for ERP-diffusion of technovation in academia-industry: A qualitative analysis. *PICMET: Portland International Center for Management of Engineering and Technology, Proceedings*, (September), 2711–2721. <https://doi.org/10.1109/PICMET.2009.5261794>
- [14] Felderer, M. (2014). Novel Methods and Technologies for Enterprise Information Systems, 8(January 2016). <https://doi.org/10.1007/978-3-319-07055-1>
- [15] Sudzina, F., & Johansson, B. (2007). Finding ERP requirements that support strategic management in organizations, (December).

- [16] Yang, H. (2016). Project Team Right-sizing for the Successful ERP Implementation. *Procedia Computer Science*, 91(I tqm), 672–676. <https://doi.org/10.1016/j.procs.2016.07.168>
- [17] Johansson, B., & de Carvalho, R. A. (2010). Software tools for requirements management in an ERP system context. *Proceedings of the 2010 ACM Symposium on Applied Computing - SAC '10*, (January), 169. <https://doi.org/10.1145/1774088.1774123>
- [18] Goud Sandhil, S., & Vishal Gupta, N. (2013). Enterprise Resource Planning (ERP) - A tool for uninterrupted supply in pharmaceutical supply chain management. *International Journal of Pharmacy and Pharmaceutical Sciences*, 5(3), 103–106.
- [19] Zamiri, A., Rostampour, A., & Nazemi, E. (2010). Towards a Holistic Requirement Management Framework for ERP Deployment. *Proceedings of the 2nd International Conference on Engineering Systems Management and Applications, ICESMA 2010*, (January), 93–98. Retrieved from <http://www.aus.edu/conferences/icesma2010/>
- [20] Rolland, C., & Prakash, N. (2001). Matching ERP system functionality to customer requirements. *Proceedings Fifth IEEE International Symposium on Requirements Engineering*, (February 2001), 66–75. <https://doi.org/10.1109/ISRE.2001.948545>
- [21] Shahid N. Bhatti, Maria Usman, Amr A. Jadi, 2015, Validation to the Requirement Elicitation Framework via Metrics. *ACM SIGSOFT Software Engineering Notes* 40(5): 17, USA. DOI= 2815021.2815031
- [22] Shahid N. Bhatti, Aneesha Rida Asghar, Atika Tabassum, “Role of Requirements Elicitation & Prioritization to Optimize Quality in Scrum Agile Development” *International Journal of Advanced Computer Science and Applications (ijacsa)*, 7(12), 2016.

Implementation of Blended Learning in Teaching at the Higher Education Institutions of Pakistan

Saira Soomro¹

Department of Distance Continuing and Computer Education,
University of Sindh, Elsa Kazi Campus, Sindh, Pakistan

Arjumand Bano Soomro²

Institute of Information & Communication Technology
University of Sindh Jamshoro, Pakistan

Tariq Bhatti³

Faculty of Education
University of Sindh, Elsa Kazi Campus,
Hyderabad-76080, Sindh, Pakistan

Najma Imtiaz Ali⁴

Department of Information Systems,
International Islamic University, Kuala Lumpur, Malaysia/
Institute of Mathematics and Computer science,
University of Sindh Jamshoro, Pakistan

Abstract—Blended learning has emerged as one of the solutions to address the various needs of Higher Education Institutions around the world. Blended Learning is the combination of traditional classroom and online endeavour. It provides advantages of both face to face learning and e-learning. The main purpose of this study is to assess the adaptation level of blended learning in teaching process at Higher Education Institutions. This study carried out mixed method approach by using explanatory sequential model. Teachers of general public universities were included as the sample for this study. Questionnaire and interview techniques were used as data gathering tools. The main findings of this study showed that teachers have a positive perception for technology usage in teaching process. Most of the teachers possessed expertise in the use of different software and equipped with internet skills. The study concluded that in blended learning implementation, universities are still at awareness level and a lot of efforts are required for effective implementation of blended learning. It is recommended that the universities' administration should provide an extra computing infrastructure (e.g. servers, bandwidth, and storage capacity) to run the courses in blended format. We recommend that in strategic plan of the universities the blended learning should be well defined and highlighted.

Keywords—Blended learning; teaching-learning; university teachers

I. INTRODUCTION

The rise of globalization has put higher education at a prominent position both in national and international context. Universities are now considered as research platforms and are playing pivotal role into their national development. Universities have also become main contributors to economic growth through the development in the field of science and technology, and through the application of modern technology. Technology has opened wide vistas of communication and digital world. Higher Education institutes are now in more challenging position to accept these revolutionary changes, so they equip the students with the new challenges of digital world. According to [1] in 21st century the use of technology became mandatory in all fields

including Higher Education and the paradigm of Higher Education has changed with penetration of technology.

In developed countries, the technological revolution has brought radical change in the field of education, particularly in higher education. The concept of blended learning has thrived in the developed world and through blended learning rapid and innovative systems in educational institutions are touching new limits and bringing advancements for knowledge seekers. The use of blended learning in combination with traditional classroom reduces the load of lecture-based teaching, and dependence on printed material. This approach of blended learning has created innovation, flexibility, activeness and collaboration in teaching-learning process. With blended learning students can use online platforms at any time and anywhere. Blended learning generated a new model for peer-to-peer interaction, peer-to-faculty interaction. Its incorporation in higher education institutions has decreased the cost of higher education [2]. Pakistan like other South Asian countries is going through many developmental changes regarding the use of technology in the field of education. In Pakistan the concept of e-Learning was first started at Allama Iqbal Open University (AIU) in year 2000. The framework was called as Open Learning Institute of Virtual Education (OLIVE). The Virtual University (VU) of Pakistan started courses via ICT, and national TV channels. Besides these two universities no other public or professional universities in Pakistan have adopted the advanced technology systems in real spirit which include blended learning [12].

The Higher Education Commission (HEC) Pakistan has also taken various initiatives to introduce and promote the use of modern technology in higher educational institutions, which includes Online Lecturing and Net-Meeting using IP-Based Video Conferencing System, Broadband Facility, National Digital Library, Pakistan Education and Research Network [12]. According to [3] the developing countries like Pakistan would not be able to get advantage from blended learning until or unless, if the factors responsible for its adaptation are persistently present in higher education institutes; e.g. ICT

penetration issues, computer literacy and hesitation to move away from traditional learning methods, electric power failures and English language barriers.

A. Rationale for the Study

The rise of Blended Learning System (BLS) has brought a paradigm shift in education and has left far-reaching impacts on higher education as well. In the developed countries the blended learning adaptation in higher education is quicker than in comparison to developing countries. This benefits the developed countries in producing trained and rich human resource from educational institutions. In case of Pakistan, the implementation of blended learning in the higher education institution is an emerging trend and facing resistance in the fully implementation in the universities. Therefore, this research carried out under title "Implementation of Blended Learning System in the Higher Education Institutions in Pakistan".

B. Research Questions

- 1) At what level blended learning is being implemented in teaching process in Higher Education Institutions?
- 2) What are the main problems & challenges faced by Higher Education Institutions in the implementation of blended learning in teaching process?

II. METHODOLOGY

This research study incorporated "Explanatory Sequential Design", a mixed method approach also called two-phase model. According to [13] this model consists of first collecting quantitative data and then collecting qualitative data to help explain or elaborate on the quantitative results provide a general picture of the research problem. The sample for the research study comprised of all four (04) public sector general universities in the province of Sindh, Pakistan. The universities included namely are; i) University of Sindh Jamshoro (UOSJ), ii) University of Karachi (UOK), iii) Shah Abdul Latif University Khairpur (SALUK) and iv) Shaheed Benazir Bhutto University Shaheed Benazirabad (SBBUSB). Two (02) departments from each Social Sciences Faculty and Natural Sciences Faculty were selected through purposive sampling technique from each university. The selection of teachers from selected departments was done through random sampling technique. A questionnaire for this research was adopted from a study by [14] on blended learning. Five-point Likert scale was used for the quantitative questionnaire starting from strongly agree to strongly disagree, whereas 1= Strongly Agree (SA), 2= agree (A), 3= undecided (UD), 4= disagree (DA) and 5= strongly disagree (SDA)

For data collection, the questionnaire was developed for university teachers, and it was administered to 58 male teachers and 27 female teachers. Out of 58 male teachers, 38 responded and out of 27 female teachers, 22 responded. Collected data was analyzed through the application of descriptive statistics (percentages, and mean score). Two teachers were also selected randomly from 04 departments of each selected university for qualitative data through interview protocol.

TABLE I. FINAL SAMPLE SIZE OF THE STUDY

Universities	Teachers at 50% for interview		Teachers at 30% for questionnaire			
	Total	50%	F	30%	M	30%
UOSJ	04	02	38	11	66	19
UOK	04	02	30	09	62	18
SALUK	04	02	15	04	38	11
SBBUSB	04	02	11	03	34	10
TOTAL sample		08		27		58

Table I shows the final sample size of respondent. There were 04 respondents for interview and the researcher chooses the 50% for the interview i.e. 2 from each university. For questionnaire the researcher choose the 30% of total respondent i.e. 11 female and 19 male from UOSJ, whereas 9 female and 18 male from UOK; furthermore 4 female and 11 male from SALUK and 3 female and 10 male from SBBUSB.

Table II shows the age of the respondent. The respondent in the range of 25-30 from all four universities were 6,4,3 and respectively, whereas respondent in the range of 31-40 were 11,12,4,and 7. Furthermore the respondents in the range of 41-50 were 7,5,4 and 2 respectively and the respondent in the range of 51- 60 from all the 4 universities were 6, 6 4 and 1 female and 11 male from SALUK and 3 female and 10 male from SBBUSB

TABLE II. AGE WISE DISTRIBUTION OF SAMPLE OF THE STUDY

Universities	Age wise distribution of teachers in years			
	25-30	31-40	41-50	51-60
UOSJ	6	11	7	6
UOK	4	12	5	6
SALUK	3	4	4	4
SBBUSB	3	7	2	1

TABLE III. DESIGNATION WISE DISTRIBUTION OF SAMPLE OF THE STUDY

Universities	Designation wise distribution of teachers			
	Lecturer	Assistant Professor	Associate Professor	Professor
UOSJ	09	12	04	05
UOK	10	10	04	03
SALUK	03	06	03	03
SBBUSB	06	05	02	00

TABLE IV. ACADEMIC QUALIFICATION WISE DISTRIBUTION OF SAMPLE

Universities	Academic qualification wise distribution of teachers		
	Master's	M.Phil	Doctorate
UOSJ	09	13	8
UOK	10	11	6
SALUK	03	07	5
SBBUSB	06	05	2

Table III shows the designation wise distribution of sample size. There were 9 lecturers 1 assistant professor, 4 associate professor and 5 professor from UOSJ. Similarly for UOK there were 10 lecturers, 10 assistant professors, 4 associate professors and 3 professors. Furthermore for SALUK and SBBUSB the respondent distribution was 3, 6, 3, 3 and 6, 5, 2 and 0 respectively.

Table IV present the academic qualification of the respondent. There were 9 masters, 13 Mphil and 8 Doctorate from UOSJ and similarly 10, 11 and 6 from UOK. Furthermore there were 3 masters 7 Mphil and 5 Doctorate from SALUK and similarly 6, 5 and 2 from SBBUSB.

III. DATA ANALYSIS

Table V shows that higher number of respondents 18.33% teachers strongly agreed, and 33.33% teachers agreed that they were advance users of Email service for teaching-learning process whereas 21.67% and 16.76% respondents were disagreed and strongly disagreed respectively. While 10.00% remain undecided. The mean score found to be 3.13. Thus, the result described that most of the teachers rated themselves as advanced users of Email service for teaching-learning process.

TABLE V. OPINION REGARDING LEVEL OF EXPERTISE IN USING EMAIL SERVICE FOR TEACHING PROCESS

Scale	SA	A	UD	DA	SDA	Mean
Frequency	11	20	06	13	10	3.13
Percentage	18.33%	33.33%	10.00%	21.67%	16.67%	

TABLE VI. OPINION REGARDING THE LEVEL OF EXPERTISE IN USING SEARCH ENGINES FOR TEACHING PROCESS

Scale	SA	A	UD	DA	SDA	Mean
Frequency	12	23	10	08	07	3.4
Percentage	20.00%	38.33%	16.67%	13.33%	11.67%	

Table VI shows the majority of teachers 20.00% strongly agreed, 38.33% agreed that they were advance users of search engines for teaching process, while 13.33% teachers disagreed and 11.67% strongly disagreed with the statement. Whereas 16.67% teachers remained undecided. The mean score found to be 3.4. Thus, the results show that most of the teachers rated themselves as advanced users of search engines for teaching in the classrooms.

TABLE VII. OPINION REGARDING LEVEL OF EXPERTISE IN USING WEB 2.0 TOOLS FOR TEACHING PROCESS

Scale	SA	A	UD	DA	SDA	Mean
Frequency	14	17	05	15	09	3.16
Percentage	23.33%	38.33%	8.33%	25.00%	15.00%	

TABLE VIII. OPINION REGARDING TEACHER'S AWARENESS OF THE BENEFITS OF BLENDED LEARNING FOR TEACHING-LEARNING PROCESS

Scale	SA	A	UD	DA	SDA	Mean
Frequency	15	18	04	10	13	3.28
Percentage	25.00%	30.00%	6.67%	16.67%	21.67%	

Table VII shows that the majority of teachers 23.33% strongly agreed, 38.33% teachers agreed that they were advance users of Web 2.0 tools for teaching process, whereas 25.00% of teachers were disagreed and 15.00% were strongly disagreed that they were basic users. While 8.33% teachers remained undecided. The mean score found to be 3.16. Thus, results show that most of the teachers rated themselves as advanced users of Web 2.0 tools for teaching.

Table VIII shows that majority of teachers with 25.00% and 30.00% were strongly agreed and agreed respectively that they were aware of the benefits of blended learning for teaching process, 16.67% of teachers disagreed, 21.67% of teachers strongly disagreed that they were unaware of the benefits of blended learning whereas 6.67% teachers remained undecided. The mean score is found to be 3.28. Thus, the results indicate that most of the teachers were aware of the benefits of blended learning for teaching process.

TABLE IX. OPINION REGARDING THE TEACHERS SUPPORT FOR BLENDED LEARNING IN TEACHING PROCESS

Scale	SA	A	UD	DA	SDA	Mean
Frequency	15	18	08	09	10	3.31
Percentage	25.00%	30.00%	13.33%	15.00%	16.67%	

TABLE X. OPINION REGARDING THE TEACHER'S VIEWS ABOUT UNIVERSITY POLICY FOR BL FOR TEACHING PROCESS

Scale	SA	A	UD	DA	SDA	Mean
Frequency	04	05	12	19	20	2.17
Percentage	6.67%	8.33%	20.00%	31.67%	33.33%	

Table IX shows that the majority of teachers 25.00% strongly agreed and 30.00% agreed that they were supporter of blended learning approach for teaching process, 15.00% of teachers disagreed with the statement whereas 16.67% strongly agreed. While 13.33% teachers remained undecided. The mean score found to be 3.31. Thus, the results show that most of the teachers were in supporter of blended learning approach in teaching.

Table X exhibits a small number of teachers 6.67% strongly agreed and 8.33% agreed with the statement that they knew about their university policy for BL for teaching

process, majority of teachers 31.67% shown their disagreed, 33.33% were strongly disagreed that they did not know about their university policy and 20.00% teachers remained undecided. The mean score is found to be 2.17. Thus, the results show that a very small number of the teachers responded that they knew their university have any policy for blended learning for teaching process.

TABLE XI. OPINION REGARDING TEACHERS' VIEWS ABOUT BLENDED LEARNING MODEL ADOPTION BY UNIVERSITY FOR TEACHING PROCESS

Scale	SA	A	UD	DA	SDA	Mean
Frequency	03	05	13	16	23	2.1
Percentage	5.00%	8.33%	21.67%	26.67%	38.33%	

Table XI shows teacher's views about Blended Learning model adoption by university for teaching process. Analysis of data exhibits a small number of teachers 5.00% were strongly agreed, 8.33% were agreed that they knew about their university model for BL for teaching-learning process, a majority of teachers with 26.67% and 38.33% ratio were disagreed and strongly disagreed respectively, and did not know about their university model while 21.67% teachers remained undecided. The mean score is 2.1. The results show that a very small number of the teachers knew that their university has any model for blended learning for teaching-learning process.

TABLE XII. OPINION ABOUT COURSE DESIGN ON BLENDED LEARNING

Scale	SA	A	UD	DA	SDA	Mean
Frequency	02	03	13	17	25	1.95
Percentage	3.33%	5.00%	21.67%	28.33%	41.67%	

Table XII shows a small number of teachers 8.33% agreed that they knew about course design on Blended Learning for teaching process, a majority of teachers 70.00% disagreed and did not know about course design and 21.67% teachers remained undecided. The mean score is found to be 1.95. The results show that a very small number of the teachers knew that their university have designed course on blended learning format for teaching-learning process.

TABLE XIII. OPINION REGARDING TEACHERS VIEWS ABOUT TRAINING RELATED TO BLENDED LEARNING COURSE DESIGN

Scale	SA	A	UD	DA	SDA	Mean
Frequency	18	24	05	06	07	3.65
Percentage	30.00%	40.00%	8.33%	10.00%	11.67%	

Table XIII shows that majority of teachers 30.00% strongly agreed, 40.00% agreed that they need trainings for blended learning course design, whereas 10.00% of teachers disagreed, 11.67% of teachers strongly disagreed that they did not want trainings for blended learning course design. While 8.33% teachers remained undecided. The mean score is found to be 3.65. The result shows that majority of teachers want trainings for designing courses on blended learning format for teaching process.

TABLE XIV. OPINION REGARDING TEACHERS VIEWS ABOUT TRAINING RELATED TO USE OF OER FOR BLENDED LEARNING COURSE

Scale	SA	A	UD	DA	SDA	Mean
Frequency	16	28	03	04	09	3.56
Percentage	26.67%	46.67%	5.00%	6.67%	15.00%	

Table XIV shows that majority of teachers with 26.67% were strongly agreed, 46.67% were agreed that they need trainings for use of Open Education Resource (OER) for blended learning course, whereas a small proportion of teachers with 6.67% were disagreed and 15.00% teachers were strongly disagreed and did not want trainings of OER for blended learning course. While 5.00% teachers remained undecided. The mean score is found to be 3.56. The result shows that majority of teachers want trainings of OER for blended learning course.

TABLE XV. OPINION REGARDING TEACHERS VIEWS ABOUT TRAINING FOR TECHNICAL STAFF FOR IMPLEMENTATION OF BLENDED LEARNING COURSE

Scale	SA	A	UD	DA	SDA	Mean
Frequency	17	26	04	06	07	3.33
Percentage	28.33%	43.33%	6.67%	10.00%	15.00%	

Table XV. shows that majority of teachers 28.33% strongly agreed, 43.33% agreed that they need trainings for technical staff for implementation of Blended Learning course, 10.00% of teachers were disagreed, 15.00% teachers were strongly disagreed did not want trainings while 6.67% teachers remained undecided. The mean score is found to be 3.33. The results show that majority of teachers want trainings .for technical staff for implementation of Blended Learning course.

TABLE XVI. OPINION REGARDING TEACHER'S VIEWS ABOUT POTENTIAL CHALLENGES OF TEACHING THROUGH BLENDED MODE TAKES MORE TIME EFFORT

Scale	SA	A	UD	DA	SDA	Mean
Frequency	16	28	03	04	09	3.56
Percentage	26.67%	46.67%	5.00%	6.67%	15.00%	

Table XVI shows that majority of teachers 26.67% strongly agreed, 46.67% agreed that teaching through blended modell is a challenge and takes more effort, 6.67% of teachers disagreed, 15.00% were strongly disagreed that they did not consider it a challenge. While 5.00% teachers remained undecided. The mean score is found to be 3.56. The results show that majority teachers consider teaching though blended mode a challenge.

TABLE XVII. OPINION REGARDING MANY TEACHERS HAVE YET TO ACCEPT THE VALUE BLENDED LEARNING

Scale	SA	A	UD	DA	SDA	Mean
Frequency	18	24	05	06	07	3.65
Percentage	30.00%	40.00%	8.33%	10.00%	11.67%	

Table XVII shows that majority of teachers 30.00% strongly agreed, 40.00% of teachers agreed that faculties needs to accept the value of teaching through blended mode. 10.00% of teachers were disagreed, 11.67% were strongly disagreed and did not consider acceptance as a challenge while 8.33% teachers remained undecided. The mean score is found to be 3.65. The results show that majority of faculties needs to accept the value of teaching through blended mode.

IV. ANALYSIS OF DATA COLLECTED FROM TEACHERS THROUGH INTERVIEWS

The interviews were conducted and analysed through discussion under the themes as follows:

A. Theme 01: Integration of Technology

Most of the participants stated that; they use computers for different activities, such as for research, lecture and presentation purpose, teachers are expert in using MS Office. Most of the teachers are expert in using e-mails and search engines for teaching-learning process, all teachers have their official email addresses given by the universities. Most of the teachers are familiar and use of web 2.0 tools like blogs and discussion forums. Most of teachers are members of professional and academic networks like academia, research gate, and LinkedIn as suggested by [4].

B. Theme 02: Level of Implementation of Blended Learning

Teachers viewed that; each individual faculty knows about the term Blended Learning, and its benefits. Even though individual faculty members are not implementing Blended Learning in their teaching process, but they advocate its need. The participants observed that faculty is not able to implement blended learning due to several reasons like lack of authorization and some exterior issues such as; no appropriate trainings of staff and teachers; and limited number of computer labs for students, no uniform definition of Blended Learning recommended officially and no office definition, no uniform Blended Learning policy is there and no strategy or no course of action, no orientation for Blended Learning, no official endorsement or no guideline for implementation of blended learning in the system, no institutional models established, or any model adopted, no courses on the blended learning format is identified in catalogues before start of any semester, there is no course is being designed yet on the blended learning format, no course is designed, which supports the Blended Learning pedagogy/instructional methods. The above same findings of understanding and examination of first phase are supported and discussed in "Blended Learning adoption and framework" by [5]. Initially, researcher assumed that University of Karachi and University of Sindh be considered at second level of early implementation phase as they both universities are urban universities and run the Directorate of Distance Education Program, a Higher Education Commission HEC, Pakistan funded project. The rest of the two universities Shah Abdul Latif University, Khairpur and Shaheed Benazir Bhutto University, Shaheed Benazirabad were assumed at first level of awareness level. But after interviews and data collection it was analyzed that all four universities of Sindh province are at same page and touches only first stage that is awareness.

C. Theme 03: Challenges

The challenges encompassed that there is no motivation and encouragement for the teachers who are using BL components into their teaching, faculty needs trainings for using OER for Blended Learning courses and faculty needs trainings for designing Blended Learning courses.

From questionnaires and interviews of teachers it is concluded that they face problems in adopting blended learning. Some of the challenges were identified as [6];

- No policy for blended learning implementation;
- No faculty support and training to initiate courses on blended format
- Lack specialized skills needed to run courses on blended format; and
- Shortage of computer laboratories to run courses on blended format.

From the participant views and questionnaire results it is clearly shown that the external factors mentioned above were affecting teachers' willpower and distracting their motivation of not opting the course on blended format. Due to lack of universities support the teachers' demands a proper system for implementation

V. DISCUSSION AND CONCLUSION

The blended learning is a new concept in Pakistan which is an underdeveloped country and has not embraced it in education. Therefore participants' trend going towards in positive direction as they are aware of it but still it's not found implemented yet. Teachers gave great importance to policies on using ICT for blended learning, and were familiar with the benefits of Blended Learning. Teachers were supporter and promoter of the Blended Learning but with no constant definition of Blended Learning, no course of action or no official approval they were at awareness level for implementation of Blended Learning. Teaching through blended mode was a time consuming, needs more efforts. Acceptation of blended instruction was a main barrier in implementation blended learning. Teachers prerequisite faculty development trainings regarding blended learning system and considered lack of funding was one of challenge for purchasing the equipment and software needed for blended learning. Moreover, teachers viewed that there was need for technical skills, assistance and support for technical problems in the implementation of blended learning system. Along with this they agreed that there was no faculty motivation or encouragement or incentive for the teachers, who were using BL components into their teaching at their own. [7] indicate that teaching blended courses can give the lecturer the opportunity to use new educational technology in the universities.[8] describes that blended learning is proficient and effective" (p. 6). A study conducted a study that explored majority of students' favoured blended learning mode because of the flexibility and convenience and Blended learning facilitate students to express the level of freedom in interacting with their peers [9].

Furthermore, they viewed that there was no model, no

course, no any official definition was there for the implementation of Blended Learning. It is also concluded that majority of teachers gave great importance to policies on using ICT for blended learning, and were familiar with the benefits of Blended Learning. There was no policy of Blended Learning was being informed by Head of Departments to concerned teachers because that there was no official approval or order, no policy, no course was being designed for the implementation of Blended Learning was planned or scheduled. According [10] in blended learning students can take advantage and proceed their learning at any time, any place at their own convenience due to its flexible nature. Blended Learning is easy in comparison to traditional face to face where time limitation and space is decided in advance and students needs to be present there if they want to learn.

It is concluded that teachers perceived that teaching through blended mode was a time consuming, needs more efforts. Acceptation of blended instruction was a main barrier in implementation blended learning. Teachers prerequisite faculty development trainings regarding blended learning system and considered lack of funding was one of challenge for purchasing the equipment and software needed for blended learning. Moreover, they viewed that there was need for technical skills, assistance and support for technical problems in the implementation of blended learning system. Along with this they agreed that there was no faculty motivation or encouragement or incentive for the teachers, who were using BL components into their teaching at their own. Furthermore, they viewed that there was no model, no course, no any official definition was there for the implementation of Blended Learning. [11] said that Blended learning is combining both potentials of face to face and online instructions. It's not the new, but a novel idea of incorporating technology with traditional methods of teaching to equip students with 21st-century skills i.e. collaboration, creativity and problem-solving skills are core areas where students expect them to become enable.

VI. RECOMMENDATIONS

This research gives few recommendations as follows:

- The university administration should provide extra computing infrastructure (servers, bandwidth, and storage capacity) to run courses in blended format.
- Universities should develop comprehensive institutional and organizational mechanisms to implement blended learning.
- In strategic plan of the universities the blended learning should be well defined and highlighted.
- In Policies and planning, the universities administration should focus on implementation of blended learning in courses.
- The Heads of the departments should also develop a comprehensive mechanism for the effective

implementation of blended learning in teaching-learning process.

- The technology-based centralized resource centre should be established to provide technical support & guidance to the teachers.
- The Learning Management System should be introduced at department level through the technology-based centralized resource centre.
- The university courses should be revised, and technological aspect must be included in the course.
- Conferences, seminar on blended learning should be organized in collaboration with virtual university Pakistan and other technology sufficient institutions.
- The separate budgetary heads should be maintained for the purchase & provision of equipment and software needed for blended learning.

REFERENCES

- [1] Young, J. R. (2002). 'Hybrid' Teaching Seeks to End the Divide between Traditional and Online Instructions. *The Chronicle of Higher Education*, 48(28), p33.
- [2] Morgan, K. R. (2002). *Blended learning: A strategic action plan for a new campus*. Seminole:University of Central Florida.
- [3] Qureshi, A &etal (2012)" Challenges of implementing e-learning in a Pakistani university" *Knowledge Management & E-Learning: An International Journal*, Vol.4, No.3, Pp 310-324.
- [4] Ndereya, N.C (2014)"Implementing Blended Learning at a Developing University: Obstacles in the way" *The Electronic Journal of e-Learning Volume 12 Issue 1 2014*, (pp101-110).
- [5] Graham, Charles, R., Woodfield, W., & Harrison, J. B. (2013). A framework for institutional adoption and implementation of blended learning in higher education. *Internet and Higher Education*. doi:10.1016/j.iheduc.2012.09.003
- [6] Alebaikan, A. R.(2010) *Perceptions of Blended Learning in Saudi Universities: Ph.D thesis*
- [7] SARAH E. KING AND KATIE CERRONE ARNOLD (2009) "Blended Learning Environments in Higher Education: A Case Study of How Professors Make it Happen" *Journal of Mid-Western Educational Researcher Volume 25, Issues ½, Pp 44-59*. Seminole, FL: University of Central Florida.
- [8] Harvey Singh. (2003). "Building Effective Blended Learning Programs, Issue of *Educational Technology*". Volume 43. Number 6, Pages 51-54.
- [9] Kistow.B (2011)" Blended learning in higher education: A study of a graduate school of business, Trinidad and Tobago" *Journal of Caribbean Teaching Scholar*,Vol. 1, No. 2, published by Educational Research Association(ERA), Pp 115–128.
- [10] Singh, Harvey. (Nov-Dec, 2003). "Building effective blended learning programs".
- [11] Garrison, D. R., & Vaughan, N. D. (2008). *Blended Learning in Higher Education: Framework, Principles, and Guidelines*. San Francisco, CA: John Wiley & Sons.
- [12] *Survey of ICTs for Education in India and South Asia, Case Studies 2010*
- [13] Creswell John W, (2012). "Educational Research Planning, Conducting, and Evaluating Quantitative and Qualitative Research", 4th Edition published by Pearson Education, Inc.
- [14] Graham, C. R., Woodfield, W., & Harrison, J. B. (2013). A framework for institutional adoption and implementation of blended learning in higher education. *The internet and higher education*, 18, 4-14.

The Measurement of Rare Plants Learning Media using *Backward Chaining* Integrated with *Context-Input-Process-Product* Evaluation Model based on Mobile Technology

Nyoman Wijana¹

Department of Biology Education
Universitas Pendidikan Ganesha
Bali, Indonesia

I Gede Astra Wesnawa³

Department of Geography Education
Universitas Pendidikan Ganesha
Bali, Indonesia

I Wayan Eka Mahendra⁵

Department of Mathematics
Education
IKIP PGRI Bali
Bali, Indonesia

Ni Nyoman Parmithi²

Department of Biology Education
IKIP PGRI Bali
Bali, Indonesia

I Made Ardana⁴

Department of Mathematics
Education
Universitas Pendidikan Ganesha
Bali, Indonesia

Dewa Gede Hendra Divayana^{*6}

Department of Information
Technology Education
Universitas Pendidikan Ganesha
Bali, Indonesia

Abstract—This research was aimed to know the effectiveness level of learning media utilization to the introduction of rare plants in *Alas Kedaton* tourism forest in *Tabanan-Bali* based on *backward chaining* for students and the general public. The type of this research includes explorative and evaluative research types. The population in this study was the plants species that exist in the *Alas Kedaton* tourism forest. The human population was the entire society in the area of *Alas Kedaton* tourism forest. The sampling method of plants species used the quadratic method, while for the human samples used purposive sampling method. The data has been collected then analyzed descriptively. The results of this study indicate that through the utilization of learning media obtained related information about the number of rare plants species in *Alas Kedaton* tourism forest as many as 48 species of plants with 26 families, and also the factors causing the scarcity of those plants species. Through the use of *CIPP (Context-Input-Process-Product)* evaluation model assisted by mobile technology, the overall average effectiveness of learning media utilization to the introduction of rare plant in *Alas Kedaton* tourism forest in *Tabanan-Bali* based on *backward chaining* amount of 88.20%, so that was included into the good categorization.

Keywords—Rare plants species; backward chaining; evaluation; CIPP; mobile technology

I. INTRODUCTION

Forests are one source of foreign exchange that has been massively exploited for timber. This exploitation leads to widespread forest loss. Until now, the destruction of the forest environment still occurs, both by the practice of illegal logging and illegal mining. Based on the data from the Planology Department of Forestry in 2010 [1], it is known that forest destruction is getting worse due to uncontrolled logging, forest fires, community utilization of timber, and the

conversion of land functions in forests. Based on data from the Bali Forestry Office in 2010 [2], the area of the mainland forest in Bali is 127,721.01 hectares or only 22.59 percent of the total area of Bali's land area of 563,286 hectares. In addition to natural disasters of drought, floods and landslides, forest destruction also causes extinction of plant species contained therein. Including local plant species that are very important for science because some of them are the types of plants that have been threatened in nature and unique plant species and endemic or have a uniqueness or very rarely found elsewhere.

Some studies may be mentioned, such as the research that has been done by Wijana in 2004 [3], 2005 [4], 2006 [5], 2008 [6]; 2009 [7], 2010 [8], 2012 [9], 2013 [10], 2014 [11], 2015 [12], and 2016 [13]-[15]. All his research is done in the area of Bali. Researches related to the analysis of terrestrial vegetation outside of Bali have been done by Arrijani, *et.al.* in 2006 [16], Sri Hartini in 2007[17], Purwaningsih in 2006 [18], Purwaningsih and Razali Yusuf in 2008 [19], Junaedi, Indrawan, and Mutaqien in 2010 [20], and Onrizal, *et.al.* in 2006 [21].

In general it can be said that the studies mentioned above, examine the composition of species, species diversity, and management of protected forests and national parks. These studies were conducted in areas such as Arrijani in Cianjur, Junaedi in West Java, Sri Hartini in East Kalimantan, Onrizal in West Kalimantan, and Purwaningsih in Southeast Sulawesi. The context of this study is more oriented to study vegetation parameters or vegetation analysis and efforts to introduce rare plants in forest areas through learning media and evaluation of the effectiveness of the utilization of learning media.

One of the forest areas used as tourist attractions in Bali is *Alas Kedaton* forest, Kukuh village, Marga district, Tabanan regency, Bali, Indonesia. According Sujaya in 2007 [22] explains that the width of *Alas Kedaton* tourism object is approximately 12 hectares, while the forest area of hedge approximately 6.5 hectares. In this forest area found the trees are large and dense, and there are several types of plants in the forest vegetation, is included in the category of rare plants.

As an effort to conserve protected forest in *Alas Kedaton* tourist area, it is necessary to introduce to society in general and the students in particular about information of rare plant species that exist in the area through mobile technology-assisted learning media. With the help of mobile technology-assisted learning media, the community and students can search and complete information about the rare plants in *Alas Kedaton* forest, Tabanan, Bali through the media whenever and wherever they are. To obtain an overview of the effectiveness of learning media utilization for introduction of rare plants in *Alas Kedaton* tourism forest, Tabanan, Bali, it is necessary to conduct an evaluation.

Generally evaluation is an activity to collect, process, and analyze a data into accurate information through a meticulous, complete and in-depth measurement process that can be useful as a recommendation for stakeholders/policy in taking a right decision. That definition of evaluation is reinforced by Divayana and Sanjaya [23], Jampel, *et.al.* [24], Arnyana, *et.al.* [25], Divayana, *et.al.* [26]-[29], Ariawan, Sanjaya, and Divayana [30], Divayana, Ardana, and Ariawan [31], Divayana [32]-[36], Sanjaya, and Divayana [37], Divayana, and Sugiharni [38], Divayana, Adiarta, and Abadi [39], Suandi, Putrayasa, and Divayana [40], Divayana, D.G.H., Adiarta, A., and Abadi [41], Sudiana, *et.al.* [42], Mahayukti, *et.al.* [43], with the core of the definition of evaluation is an activity to obtain recommendations so that it can be used as a basis for decision-making to continue/stop the program being evaluated.

There are several evaluation models that can be used in an evaluation such as: Goal Free Evaluation Model, Goal Oriented Evaluation Model, Responsive Evaluation Model, Formative-Summative Evaluation Model, Countenance Evaluation Model, Center for the Study of Evaluation-University of California in Los Angeles, CIPP (Context, Input, Process, Product), and Discrepancy Model.

From some of these models, the most suitable and appropriate model used in this study is the *CIPP* evaluation model, because this model can provide related information: 1) the evaluation *context* that provides value and description of the things that cause learning media to introduce of rare plants in *Alas Kedaton* tourism forest can be realized, 2) evaluation inputs that determine the available resources, alternative strategies and what plans should be done to encourage the holding of learning media, 3) evaluation process that provides value and description of the activities that have been implemented to achieve the objectives of the implementation of the learning media, and 4) evaluation products that provide value and description of the results achieved after utilizing the learning media.

From the description above, the problems studied in this research are: 1) How the use of learning media to introduce of rare plants, especially in *Alas Kedaton* tourism forest to know the number of species of rare plants in that forest and the factors causing the scarcity of the plant species; 2) What is the effectiveness level of utilization of learning media to introduce rare plants in *Alas Kedaton* tourism forest for students and the general public?

Based on the problems and the use of a new innovation in the form of learning media as a solution to problem solves the existing problems, so the researchers are interested in conducting research studies about the effectiveness measurement of the learning media for introduction of rare plants in *Alas Kedaton* tourism forest in Tabanan-Bali using *backward chaining* integrated with *Context-Input-Process-Product* evaluation model based on mobile technology

II. RESEARCH METHODOLOGY

The type of this research includes explorative research and evaluative research. It said explorative research because it explores of rare plant species in the *Alas Kedaton* forest tourism in Tabanan, Bali, Indonesia. It is said evaluative research for evaluating of learning media to introduce of rare plants in *Alas Kedaton* forest tourism. The explorative research location in *Alas Kedaton* tourism forest is with an area of 6.5 hectares. While the location of evaluative research conducted in the area of *Alas Kedaton* tourism object and high school around in *Alas Kedaton* Tabanan.

Population in this explorative research was plant species that exist in *Alas Kedaton* tourism forest. The population of evaluative research was the entire community in the area of *Alas Kedaton* tourism forest. The sampling method of plant species for explorative research was using the quadratic method [11], [44], while the community sampling method for evaluative research is by using purposive Sampling. The samples of plant species are all plant species covered by squares of 20 x 20 m size as many as 100 squares. For the sample of the community was taken as many as 25 people.

In the sampling technique of plant species using systematic squares, the squares are placed continuously at 10 x 20m intervals along the line of the compass line, as many as 100 squares. Each square is recorded for its constituent plant species. Plant species that have been collected then determined the species of plants that fall into the rare category. The determination of this rare plant species is done by studying existing documents, conducting interviews, and seeking information from various sources. Furthermore, with in-depth interviews with sources of informants from the community around the forest area, and including the District and Provincial Forest Service, to obtain information related to rare plants that fall into the national rare category, rare at the level of Bali province, scarce at Tabanan regency level, and Rare at Marga and Kukuh Village levels. Further data were analyzed descriptively. In purposive sampling technique of society in evaluating learning media conducted with the intention of involving parties who have interests/goals and understand the object/program studied in this case related to learning media to introduce of rare plants in *Alas Kedaton* forest tourism. The evaluation results using the *CIPP* model

on the use of learning media to introduce the rare plants in the *Alas Kedaton* tourism forest are indicated by the average percentage of effectiveness calculated using the following percentage descriptive formula [45].

$$\Sigma(\text{Answer} * \text{Weight of Each Choice})$$

$$\text{Percentage} = \frac{\text{---}}{n * \text{The Highest Weight}} * 100\% \quad (1)$$

Notes:

$$\Sigma = \text{Amount}$$

$$n = \text{Total number of questionnaire items}$$

Furthermore, to calculate the percentage of all subjects used by using the following formula:

$$F$$

$$\text{Percentage} = \frac{\text{---}}{N} \quad (2)$$

Notes:

F = Total percentage of the entire subject

N = Number of subjects

To be able to give meaning and decision on the percentage level of effectiveness/achievement, then used scale conversion effectiveness level as follows [45] (Table I):

TABLE I. CONVERSION LEVEL EFFECTIVENESS BY SCALE OF FIVE

Level of Effectiveness	Category
90-100 %	Very Good
80-89 %	Good
65-79 %	Enough
55-64 %	Less
0-54 %	Very Less

III. RESULTS AND DISCUSSION

A. Result

Recapitulation of explorative research results on plant species present in *Alas Kedaton* tourism forest, presented in detail in the Table II.

There are a total of 48 plant species found in the *Alas Kedaton* tourism forest, which belongs to 26 families, with details of the following families: *Meliaceae* (8 species), *Moraceae* (7 species), *Lauraceae* (3 species), *Annonaceae* (3 species) (2 species), *Apocynaceae* (2 species), *Sterculiaceae* (1 species), *Lythraceae* (1 species), *Euphorbiaceae* (1 species), *Clusiaceae* (2 species), *Myocycaceae* (2 species) (1 species), *Phyllanthaceae* (1 species), *Rubiaceae* (1 species), *Caesalpiniaceae* (1 species), *Sabiaceae* (1 species), *Elaeocarpaceae* (1 species), *Verbenaceae* (1 species), *Malpighiaceae* (1 species), *Cornaceae* (1 species), *Rubiaceae* (1 species), and *Leeaceae* (1 species). From the floristic list of plants above, then by using literature/document review, interviews, and some relevant information, a rare plant species is obtained as presented in Table III.

TABLE II. LIST OF FLORISTIC SPECIES OF COMMON SPECIES IN ALAS KEDATON TOURISM FOREST TABANAN, BALI, INDONESIA

No	Family	Name of Plant Species	
		Local Name*/Indonesia	Scientific Name
1.	<i>Anacardiaceae</i>	Dau	<i>Dracontomelum mangiferum</i>
		Mete Mini	<i>Semecarpus cassuvium</i>
2.	<i>Annonaceae</i>	Sandat	<i>Cananga odorata</i>
		Blakatak	<i>Polyalthia lateriflora</i>
		Kayu Madas	<i>Polyalthia korinti</i>
3.	<i>Apocynaceae</i>	Pulai/Pule	<i>Alstonia scholaris</i>
		Bukak	<i>Rauwolfia javanica</i>
4.	<i>Arecaceae</i>	Rotan	<i>Calamus axillaris</i>
		Jaka/Aren	<i>Arenga pinnata</i>
5.	<i>Caesalpiniaceae</i>	Benul	<i>Parkia speciosa</i>
6.	<i>Clusiaceae</i>	Badung	<i>Garcinia divica</i>
7.	<i>Combretaceae</i>	Kayu Kunyit	<i>Terminalia sumatrana</i>
8.	<i>Cornaceae</i>	Jelit-jelit	<i>Alangium salviifolium</i>
9.	<i>Elaeocarpaceae</i>	Genitri	<i>Elaeocarpus ganitrus</i>
10.	<i>Euphorbiaceae</i>	Buni Hutan	<i>Antidesma bunius</i>
		Bejulitan	<i>Litsea glutinosa</i>
		Kayu Besi	<i>Eusideroxylon zwageri</i>
11.	<i>Lauraceae</i>	Kayu Manis	<i>Cinnamomum burmani</i>
		Putat/ Kutat	<i>Planchonia valida</i>
12.	<i>Lecythidaceae</i>	Gegirang	<i>Leea sp.</i>
13.	<i>Lythraceae</i>	Tangi/Bungur	<i>Lagerstroemia speciosa</i>
14.	<i>Malpighiaceae</i>	Bergiding	<i>Hiptage benghalensis</i>
		Majegau	<i>Dysoxylum densiflorum</i>
15.	<i>Meliaceae</i>	Kayu Adeng	<i>Dysoxylum caulostachyum</i>
		Kepohpoh	<i>Buchanania arborescens</i>
		Kayu Bawang	<i>Dysoxylum alliaceum</i>
		Kayu Nyoling	<i>Pisnoid umbellata</i>
		Sentul	<i>Sandoricum koetjape</i>
		Mahoni	<i>Swietenia mahagoni</i>
16.	<i>Moraceae</i>	Langsat Lutung	<i>Aglaiia argentea</i>
		Beringin Hijau	<i>Ficus benyamina</i>
		Teep/Terep	<i>Artocarpus elastica</i>
		Ae/ Ara	<i>Ficus racemosa</i>
		Bunut	<i>Ficus altissima</i>
		Serut/Pungut	<i>Streblus asper</i>
		Kacu-Kacu	<i>Ficus magnoliaefolia</i>
Awar-Awar	<i>Ficus septica</i>		
17.	<i>Myrsinaceae</i>	Lampeni	<i>Ardisia humilis</i>
18.	<i>Myristicaceae</i>	Kayu Anak	<i>Knema laurina</i>
19.	<i>Myrtaceae</i>	Kaliampuak/ Jambu Hutan	<i>Eugenia densiflora</i>
		Salam	<i>Syzygium polyanthum</i>
20.	<i>Phyllanthaceae</i>	Gintungan	<i>Bischofia javanica</i>
21.	<i>Rubiaceae</i>	Kayu Nyan-Nyan	<i>Guettarda speciosa</i>
		Jarum-Jarum	<i>Pavetta subvelutina</i>
22.	<i>Sabiaceae</i>	Kayu Sambuk	<i>Meliosma pinnata</i>
23.	<i>Sapotaceae</i>	Nyantuh	<i>Palaquium javanicum</i>
24.	<i>Sterculiaceae</i>	Bayur	<i>Pterospermum javanicum</i>
		Kayu Taluh	<i>Vitex glabrata</i>

Source: Wijana in 2018 [46], Wijana and Setiawan in 2017 [47] Notes: *) Local Name Using Balinese Language

TABLE III. LIST OF RARE PLANTS SPECIES IN ALAS KEDATON TOURISM FOREST, TABANAN, BALI, INDONESIA

No.	Family	Name of Plant Species		Number of Individuals	Status
		Local Name*) /Indonesia	Scientific Name		
1.	Anacardiaceae	Dau	<i>Dracontomelum mangiferum</i>	8	BR
		Mete Mini	<i>Semecarpus cassuvium</i>	1	BR
2.	Annonaceae	Sandat	<i>Cananga adorata</i>	2	NR
		Blakatak	<i>Polyalthia lateriflora</i>	7	TR
		Kayu Madas	<i>Polyalthia korinti</i>	17	MR
3.	Apocynaceae	Pulai/ Pule	<i>Alstonia scholaris</i>	1	NR
		Bukak	<i>Rauwolfia javanica</i>	78	TR
4.	Arecaceae	Rotan	<i>Calamus axillaris</i>	6	BR
		Jaka/ Aren	<i>Arenga pinnata</i>	2	BR
5.	Caesalpinioidea	Benul	<i>Parkia speciosa</i>	4	BR
6.	Clusiaceae	Badung	<i>Garcinia divica</i>	1	NR
7.	Combretaceae	Kayu Kunyi	<i>Terminalia sumatrana</i>	9	BR
8.	Elaeocarpaceae	Genitri	<i>Elaeocarpus ganitrus</i>	1	BR
9.	Euphorbiaceae	Buni Hutan	<i>Antidesma bunius</i>	2	NR
10.	Lauraceae	Bejulitan	<i>Litsea glutinosa</i>	26	BR
		Kayu Besi	<i>Eusideroxylon zwageri</i>	7	BR
		Kayu Manis	<i>Cinnamomum burmani</i>	57	TR
11.	Lecythidaceae	Putat/ Kutat	<i>Planchonia valida</i>	12	BR
12.	Lythraceae	Tangi/Bungur	<i>Lagerstroemia speciosa</i>	9	NR
13.	Malpighiales	Bergiding	<i>Hiptage benghalensis</i>	79	TR
14.	Meliaceae	Majegau	<i>Dysoxylum densiflorum</i>	5	NR
		Kayu Adeng	<i>Dysoxylum caulostachyum</i>	23	BR
		Kepohpoh	<i>Buchanania arborescens</i>	10	BR
		Kayu Bawang	<i>Dysoxylum alliaceum</i>	60	TR
		Kayu Nyoling	<i>Pisnoid umbellata</i>	4	TR
		Sentul	<i>Sandoricum koetjape</i>	3	TR
		Mahoni	<i>Swietenia mahagoni</i>	63	MR
Langsat Lutung	<i>Aglaia argentea</i>	13	MR		
15.	Moraceae	Beringin Hijau	<i>Ficus benyamina</i>	1	NR
		Teep/Terep	<i>Artocarpus elastic</i>	32	BR
		Ae/Ara	<i>Ficus racemosa</i>	18	BR
		Bunut	<i>Ficus altissima</i>	2	BR
		Serut/Pungut	<i>Streblus asper</i>	2	TR
Kacu-Kacu	<i>Ficus magnoliaefolia</i>	5	MR		
16.	Myristicaceae	Kayu Anak	<i>Knema laurina</i>	5	BR
17.	Myrtaceae	Kaliampuak/ Jambu Hutan	<i>Eugenia densiflora</i>	11	TR
18.	Phyllanthaceae	Gintungan	<i>Bischofia javanica</i>	5	BR
19.	Rubiaceae	Kayu Nyan-Nyan	<i>Guettarda speciosa</i>	4	BR
20.	Sabiaceae	Kayu Sambuk	<i>Meliosma pinnata</i>	3	BR
21.	Sapotaceae	Nyantuh	<i>Palaquium javanicum</i>	34	BR
22.	Sterculiaceae	Bayur	<i>Pterospermum javanicum</i>	11	NR
23.	Verbenaceae	Kayu Taluh	<i>Vitex glabrata</i>	1,275	TR

Source: Wijana in 2018 [46], Wijana and Setiawan in 2017 [47]

TABLE IV. SOME EXAMPLES OF COMPLETE INFORMATION ABOUT RARE PLANT SPECIES IN ALAS KEDATON TOURISM FOREST

1. Kayu Taluh (<i>Vitex glabrata</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Lamiales</i> Family : <i>Verbenaceae</i> Genus : <i>Vitex</i> Species : <i>Vitex glabrata</i>
	The plant is a tree, its height reaches \pm 25 m, stem diameter 35 - 45 cm, this tree has many branch which is not straight/bent and irregular. The wood is quite hard, solid, the fiber is straight, the color is greenish to yellow brown. The leaves pinnate with the shape of the round leaves of eggs until tapering/ellipse and tapered to the tip and base of the leaves.
2. Kayu Bawang (<i>Dysoxylum alliaceum</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Sapindales</i> Family : <i>Meliaceae</i> Genus : <i>Dysoxylum</i> Species : <i>Dysoxylum alliaceum</i>
	Plants with height up to 20-25 m in diameter of stems 40-60 cm. The trunk is straight with white wood without a terrace. The leaves are pinnate with a sitting leaf opposite the shape of the lanceolate.
3. Tangi/Bungur (<i>Lagerstroemia speciosa</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Myrtales</i> Family : <i>Lythraceae</i> Genus : <i>Lagerstroemia</i> Species : <i>Lagerstroemia speciosa</i>
	Plants with a height of 10-30 m. Round stem, branching starting from the base, light brown. Leaves single, stiff, short stem. The leaves are oval shaped, elliptical, with a length of 9-28 cm and dark green. Compound interest, arranged in panicles. The fruit is a box, ball-shaped until rounded elongated, with a length of 2-3.5 cm, has a space as much as 3-7, fruit is still young green, gradually become brown.
4. Kayu Besi (<i>Eusideroxylon zwageri</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Ranales</i> Family : <i>Lauraceae</i> Genus : <i>Eusideroxylon</i> Species : <i>Eusideroxylon zwageri</i>
	Plants with a height of 10 m. The trunk is strong but the shape of the trunk is bent. Leaves pinnate, pointed leaf tip, rounded base of leaf, flat leaf edge. Twigs are reddish brown. The fruit of this plant is a fruit stone, shaped ellipse to make, seed one with a length of 7-16 cm and width 5-9 cm.
5. Kayu Jelemal/Kayu Anak (<i>Knema laurina</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Magnoliales</i> Family : <i>Myristicaceae</i> Genus : <i>Knema</i> Species : <i>Knema laurina</i>
	Plants with a height of \pm 20 m. The trunk is light brown with red sap. Has an <i>arillus</i> that covers all the pink seeds. Leaf blade pinnate with a lanceolate shape, with a slippery leaf surface.
6. Majegau (<i>Dysoxylum densiflorum</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Sapindales</i> Family : <i>Meliaceae</i> Genus : <i>Dysoxylum</i> Species : <i>Dysoxylum densiflorum</i>
	Plants with a height of 40 m with a diameter of 1.2 m. The trunk is woody, the wood is heavy, hard but fibrous with a light brown to pink or brown-pink, shiny. <i>Majegau</i> leaves are oval shaped lancet. The fruit is oval-shaped with a length of between 3-6 cm of brown to orange.
7. Gintungan (<i>Bischofia javanica</i>)	

	<p>Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Malpighiales</i> Family : <i>Phyllanthaceae</i> Genus : <i>Bischofia</i> Species : <i>Bischofia javanica</i></p> <p>Plants with a height of ± 40 m, stem diameter 95 - 150 cm. The trunk is straight, no wood or <i>bomi</i> root, no grooved. Shape of round leaves of eggs that share/notched three and tapered to the tip of the leaf. Seated leaf or spiral/circular location, has a long leaf stalk. Inflorescences of the shape of the panicle, small, located at the end of the stem with a long flower stalk. The fruit is also small (1.2 - 1.5 cm).</p>
<p>8. Sentul (<i>Sandoricum koetjape</i>)</p>	
	<p>Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Sapindales</i> Family : <i>Meliaceae</i> Genus : <i>Sandoricum</i> Species : <i>Sandoricum koetjape</i></p> <p>Plants with a height of 30 m, with a diameter of 90 cm, gummy like milk. Compound leaves alternate, pinnate with three leaves, rounded or slightly pointy shape at the base, tapering at the end; Shiny green on the top, a dull green beneath it. Flowers in panicles in the armpits of leaves, haired, dangling, up to 25 cm. The fruits of <i>buni</i> are rounded slightly flat, 5-6 cm, yellow or reddish if ripe, fluffy like velvet.</p>
<p>9. Bunut (<i>Ficus altissima</i>)</p>	
	<p>Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Urticales</i> Family : <i>Moraceae</i> Genus : <i>Ficus</i> Species : <i>Ficus altissima</i></p> <p>Plants with a height of 20-30 m. Stems woody, cylindrical, dark brown, smooth surface, branches spread irregularly to form a shady tree, out roots hanging from the trunk or branch that has been large. Single leaves, stemmed, arranged alternately, elliptic, tapered ends (<i>acuminatus</i>), flat edges, shiny surfaces (<i>nitidus</i>), and have slippery leaf surfaces.</p>
<p>10. Pulai/Pule (<i>Alstonia scholaris</i>)</p>	
	<p>Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Gentianales</i> Family : <i>Apocynaceae</i> Genus : <i>Alstonia</i> Species : <i>Alstonia scholaris</i></p> <p>Plants with a height of 10-50 m. The trunk is straight, straight, dark green. Single leaf, shaped lanceolate, rounded edges and tapered, flat edge. Stained white and sticky, bone leaves tightly, circular center leaves 4-8 strands. The flowers are compound, the shape is panicle, with the oval petals. The fruit is ribbon-shaped with a length of 20-50 mm and is white.</p>
<p>11. Genitri (<i>Elaeocarpus ganitrus</i>)</p>	
	<p>Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Malvales</i> Family : <i>Elaeocarpaceae</i> Genus : <i>Elaeocarpus</i> Species : <i>Elaeocarpus ganitrus</i></p> <p>Plants with a height of 20-30 m. The stems are erect, woody, round, and rough-colored brown. Leaves single, green, oval-shaped with serrated edge, tip and base tapered, long= 8-20 cm and width= 3-6 cm. Flower type is flowers compound shaped panicles. Jenitri fruit are <i>buni</i> type, round, green. The seeds are round, brown to dark brown in diameter between 0.5 cm - 2 cm. The surface of the hollow and grooved (threaded) seeds are carved.</p>
<p>12. Badung/Mundu (<i>Garcinia divica</i>)</p>	
	<p>Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Malpighiales</i> Family : <i>Clusiaceae</i> Genus : <i>Garcinia</i> Species : <i>Garcinia divica</i></p> <p>Plants with height of 13-15 m. The trunk has brown leather and white gummy. The leaves are oval-shaped to oval with a length of 10-30 cm. The flowers are whitish yellow.</p>
<p>13. Nyantuh (<i>Palaquium javanicum</i>)</p>	

	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Ericales</i> Family : <i>Sapotaceae</i> Genus : <i>Palaquium</i> Species : <i>Palaquium javanicum</i>
Plants with height of up to 30 m and a diameter of 0.5 m. Trunked upright with brownish red. The bark is yellow to red, and the sap is white. Leaf single with round breech shape. Flower lops on leaf axillary.	
14. Teep/Terep (<i>Artocarpus elastica</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Urticales</i> Family : <i>Moraceae</i> Genus : <i>Artocarpus</i> Species : <i>Artocarpus elastica</i>
Plants with height of 25 m with a trunk diameter of up to 80 cm. The leaves are large with a length of up to ± 50 cm, single, pinnate, upper and lower leaf surfaces furry so coarse textured. The fruit is compound and is protected with the skin of a prickly soft fruit.	
15. Putat/Kutat (<i>Planchonia valida</i>)	
	Kingdom : <i>Plantae</i> Division : <i>Magnoliophyta</i> Class : <i>Magnoliopsida</i> Order : <i>Lecythidales</i> Family : <i>Lecythidaceae</i> Genus : <i>Planchonia</i> Species : <i>Planchonia valida</i>
Plants with a height of up to 50 m, diameter of 200 cm, with stems upright, straight, and watery. The headboard is round, bushy, dark green and shiny, which in the dry season leaves fall and before the autumn leaves red. The bark is grayish brown to dark brown, peeling off in the form of small pieces. Inflorescence in the form of bunches. The flowers have many stamens. The fruit is oval.	

Notes:

NR: National Rare is protected by law [48]

BR: Rare in Bali is protected by law [48]

TR: Rare in Tabanan regency

MR : Rare in Marga sub-district

*) Local Name Using Balinese Language

From the 48 plant species commonly found in the *Alas Kedaton* tourism forest, there are as many as 42 (87.5%) plant species belonging to the rare category. This rare plant category is based on document/literature studies with reference to the Forest Service which has established several rare plant species. In addition it is also based on interviews with sources of informants around the *Alas Kedaton* forest tourism; also, accompanied by interviews to people who generally live outside the *Alas Kedaton* forest tourism, even to people outside Tabanan regency. From the results of literature studies and interviews with communities and Provincial and District Forestry Offices, rare plant categories such as rare national species, scarce at the level of the Bali province, scarce at level of Tabanan district, and rare at Marga and Kukuh sub-districts. From Table III, it appears that there are 8 (19.04%) plant species belonging to the national rare category, 20 (47.62%) of rare plant species at Bali Province level, 10 (23.81%) of rare plant species at Tabanan regency level, and 4 (9.52%) species belonging to the rare category at Marga sub-district level, including rare in the Kukuh village level. When viewed from the number of individual species that exist, from the square of 20x20m as many as 100 squares obtained species of rare plants with the largest number of individuals is plant species of Kayu Taluh (*Vitex glabrata*), with an individual number of 1,275 individuals. While the least number of

individual species are: Beringin Hijau (*Ficus benyamina*), Pulai/Pule (*Alstonia scholaris*), Badung (*Garcinia divica*), Mete Mini (*Semecarpus cassuvium*, and Genitri (*Elaeocarpus ganitrus*), with an each individual number of 1 individual. Thus it can be stated that in the *Alas Kedaton* forest tourism, as a place of conservation of rare plants, because quite a lot of rare plant species that exist in the forest. It also appears that the number of individuals belonging to the rare plant category is found to be only one individual species in size 20 × 20 m × 100 m with the interval spacing of 10 × 20 m; so very apprehensive for plant species with such conditions. This needs special attention for local tourism forest managers. Below are some examples of rare plant species present in *Alas Kedaton* tourism forest, Tabanan, Bali, Indonesia. Some examples of complete information about rare plant species found in *Alas Kedaton* tourism forest, Tabanan, Bali, Indonesia can be seen in Table IV.

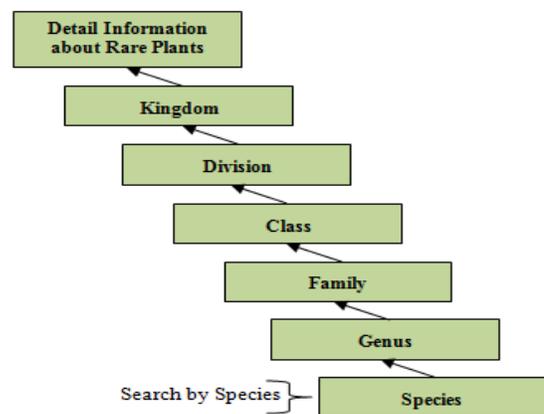


Fig. 1. The Search Process of Detail Information about Rare Plants by Species Name Using Backward Chaining Concept.



Fig. 2. The Display of Learning Media to Introduce of Rare Plants in *Alas Kedaton* Tourism Forest, Tabanan, Bali Based on *Backward Chaining* Concept.

Data from the explorative research results about complete information of rare plant species in *Alas Kedaton* tourism forest, then used as a knowledge base that is incorporated into the learning media, while the concept of *backward chaining* is used in searching for complete information about rare plant species. The view of the use of the concept of *backward chaining* and display of learning media to introduce of rare plants in *Alas Kedaton* tourism forest, Tabanan, and Bali can be seen in Fig. 1 and 2.

The standard used to measure the effectiveness of use of the media learning to introduce the rare plants in terms of *CIPP* evaluation model components can be seen in Table V.

TABLE V. MEASUREMENT STANDARD OF EFFECTIVENESS OF USE OF MEDIA LEARNING TO INTRODUCE OF RARE PLANTS IN TERMS OF CIPP EVALUATION MODEL COMPONENTS

No	Evaluation Components	Evaluation Aspects	Measurement Standard of Effectiveness (%)
1.	Context	A ₁ Aim	88
		A ₂ Legality	90
		A ₃ Stakeholders Support	85
2.	Input	A ₄ Facilities and infrastructure	80
		A ₅ Knowledge Base	88
		A ₆ Human Resources	80
		A ₇ Funding	80
3.	Process	A ₈ The ability of development team to manage the rule	85
		A ₉ The ability of development team to manage the knowledge base	90
		A ₁₀ The ability of development team to package/ present the media to be interactive	85
		A ₁₁ The ability of users in using information technology	80
4.	Product	A ₁₂ Interactivity of media	85
		A ₁₃ Accuracy of information	88
		A ₁₄ Easy access	85
		A ₁₅ Display of media design	85

Table V above shows the scores of measurement standard of effectiveness that was used as a basic reference in deciding evaluation. If the measurement results of the learning media to introduce of rare plants in *Alas Kedaton* tourism forest, Tabanan, Bali shows a less value than the scores of measurement standard of effectiveness, so that the application can be said to be ineffective while if the value is equal or even exceeds of the standard, then the application can be said to be effective. The effectiveness measurement results of the utilization of learning media to introduce of rare plants in *Alas Kedaton* tourism forest, Tabanan, Bali using *CIPP* evaluation model can be seen in Table VI.

Table VI above shows the effectiveness measurement results of the use of learning media to introduce of rare plants in *Alas Kedaton* tourism forest, Tabanan, Bali using *CIPP* evaluation model, conducted by 25 respondents with giving an assessment of 15 evaluation aspects. Percentage of effectiveness in aspect-1 (*aim*) was 89.60%, in aspect-2 (*legality*) was 92.00%, in aspect-3 (*stakeholders support*) was 85.60%, in aspect-4 (*facilities and infrastructure*) was 83.20%, in aspect-5 (*knowledge base*) was 89.60%, in aspect-6 (*human resources*) was 81.60%, in aspect-7 (*funding*) was 82.40%, in aspect-8 (*the ability of development team to manage the rule*) sebesar 88.80%, in aspect-9 (*the ability of development team to manage the knowledge base*) was 92.00%, in aspect-10 (*the ability of development team to package/present the media to be interactive*) was 89.60%, in aspect-11 (*the ability of users in using information technology*) was 80.80%, in aspect-12 (*interactivity of media*) was 88.00%, in aspect-13 (*accuracy of information*) was 91.20%, in aspect-14 (*easy access*) was 87.20%, and in aspect-15 (*display of media design*) was 86.40%.

TABLE VI. THE MEASUREMENT RESULTS OF EFFECTIVENESS OF THE USE OF LEARNING MEDIA TO INTRODUCE RARE PLANTS IN ALAS KEDATON TOURISM FOREST, TABANAN, BALI USING CIPP EVALUATION MODEL

No	Respondent	CIPP Evaluation Component														
		Context			Input				Process				Product			
		Context Aspects			Input Aspects				Process Aspects				Product Aspects			
		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
1	R ₁	5	5	4	4	5	4	4	4	5	5	4	4	5	4	4
2	R ₂	4	5	4	5	4	4	4	5	5	4	4	5	5	4	4
3	R ₃	4	4	5	4	5	5	4	4	5	5	3	4	5	5	4
4	R ₄	5	4	4	4	4	4	4	4	4	5	5	5	4	4	5
5	R ₅	5	5	4	4	4	5	4	4	5	5	5	4	4	4	4
6	R ₆	4	4	4	4	5	4	5	4	5	5	3	5	5	5	4
7	R ₇	5	5	4	4	4	4	4	5	4	4	4	4	4	4	4
8	R ₈	4	4	5	4	5	4	4	5	5	5	4	5	5	5	5
9	R ₉	4	5	4	5	4	4	4	4	5	4	3	4	4	4	4
10	R ₁₀	4	5	4	4	4	5	4	5	4	5	4	4	5	4	4
11	R ₁₁	4	5	4	4	4	4	4	4	4	4	4	5	4	5	4
12	R ₁₂	4	5	5	4	5	3	4	5	4	5	4	4	4	4	4
13	R ₁₃	5	4	4	4	5	4	4	4	5	4	5	5	5	5	5
14	R ₁₄	5	5	5	4	5	3	4	4	5	4	4	4	4	4	4
15	R ₁₅	5	4	4	4	4	4	4	5	4	4	5	4	5	4	5
16	R ₁₆	4	5	5	5	5	3	5	4	5	5	4	5	4	5	5
17	R ₁₇	5	4	4	4	4	4	4	5	4	4	3	4	5	4	4
18	R ₁₈	4	5	4	4	5	4	4	4	5	4	4	5	4	5	4
19	R ₁₉	5	4	4	4	4	4	4	4	4	4	4	4	5	4	5
20	R ₂₀	4	5	5	4	5	3	4	5	5	4	5	4	5	5	4
21	R ₂₁	4	5	4	5	4	4	4	4	4	5	5	4	4	4	4
22	R ₂₂	5	4	4	4	5	5	4	5	5	4	3	5	5	4	5
23	R ₂₃	5	5	5	4	4	4	5	5	5	4	4	4	4	5	4
24	R ₂₄	4	5	4	4	4	5	4	4	4	5	4	5	5	4	4
25	R ₂₅	5	4	4	4	5	5	4	5	5	5	4	4	5	4	5
Total		112	115	107	104	112	102	103	111	115	112	101	110	114	109	108
Percentage of Effectiveness Each Aspect (%)		89.60	92.00	85.60	83.20	89.60	81.60	82.40	88.80	92.00	89.60	80.80	88.00	91.20	87.20	86.40
Percentage of Effectiveness Each Component (%)		89.07			84.40				90.13				88.80			
Average of Overall Components (%)		88.20														

Based on the percentage of effectiveness in each aspect, so the percentage of effectiveness on *Context* components can be determined by the amount of 89.07% (including the effectiveness level in the good category). The Input Component was 84.40% (including the effectiveness level in the good category). The Process component was 90.13% (including the effectiveness level in the good category). The Product component was 88.80% (including the effectiveness level in the good category). The measurement results of the effectiveness of the use of learning media to introduce of rare plants in *Alas Kedaton* tourism forest, Tabanan, Bali using CIPP evaluation model based on mobile technology can be seen in Fig. 3.

Based from Table III it is clear that there are as many as 42 (87.5%) rare plant species from a total of 48 plant species present in the *Alas Kedaton* tourism forest. Meanwhile, according to the Provincial Forestry Office of Bali in 1987 from about 200 rare plants in Indonesia which IUCN category

(International Union for Conservation of Nature) in 1987, as many as 32 plants are already known in Bali. The amount of vegetation/flora in the *Alas Kedaton* tourism forest conducted in 2003 and 2005, the type of plants at this time experiencing a change that increases. In 2003 and 2005, 29 species of rare plants from 43 plant species were identified, while 42 species of rare plants from 46 plant species were identified.

That change is influenced by various factors from the environment and the activity of living things in it. Indriyanto in 2006 [49] explained that community of plants have dynamics or changes, both caused by the activity of nature and humans. Sugita in 2015 [50] explains that changes in the natural environment or the composition of plants in a region can be caused by adaptation to soil environment, topography, geology and climate conditions, through changes in body and function, while the environment also undergoes changes through physical or biogeochemical processes to maintain quality Life support and balance of community systems.

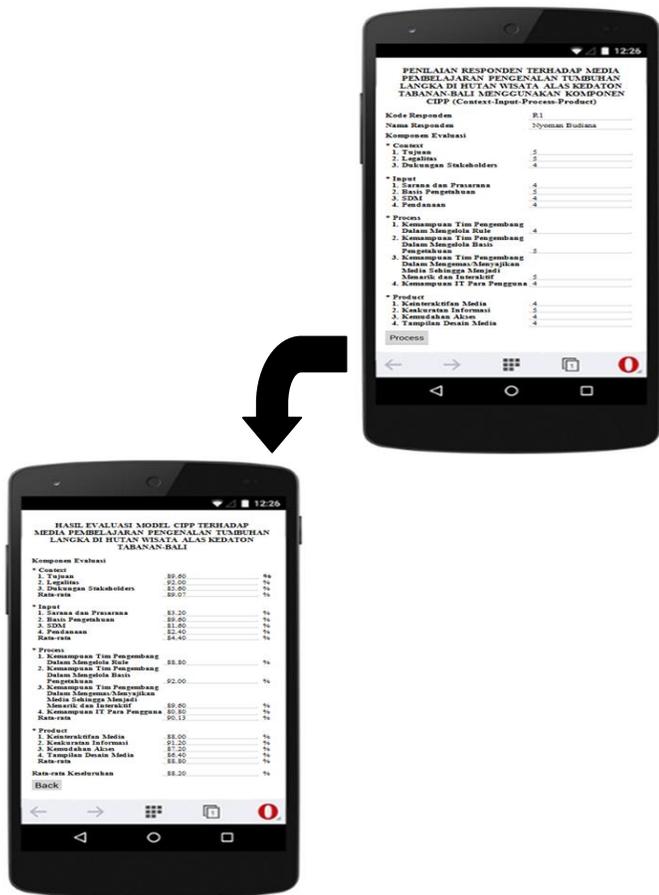


Fig. 3. The Display of Measurement Results of the use of Rare Plants Learning Media using CIPP Evaluation Model Based on Mobile Technology.

B. Discussion

The statement is in accordance with the results of interviews with the manager of *Alas Kedaton* tourism forest explained that changes in the composition of plants that have changed the occurrence of the number of rare plants that exist today, due to the efforts of planting plants in forest tours by the spread of new plant seeds. On the other hand, the existence of rare plants in the forest area, is old and dead, and not accompanied by replanting. In addition, according to the forest manager, some plants also died due to the influence of animal disturbance in this forest area, especially animal group "*Pteropus vampyrus*" which occupy the plant "*Pterospermum javanicum*" as its habitat, thus disrupting the growth of the tree.

There are several opinions that suggest a plant species may become scarce. The factors that causing plants to become scarce, can be grouped as follows: 1) Rare naturally as a result of a-biotic factors (fire, drought) or biotic (pest or disease). This naturally occurring scarcity process is especially susceptible to endemic plant species that are clustered in certain areas such as the *Rafflesia arnoldi* plant in West Sumatra or non-endemic plants but relatively small populations and very distant population distribution such as *Sawo Kecil* (*Manilkara kauki*) plant in Blambangan Jawa East, West Bali Grand Prapat, and Pedan in Sumbawa. Theoretically, the loss or scarcity of a species will affect the

survival of other co-evolutionary species [51]; 2) Rare as a result of human actions directly or indirectly. It can directly be an excessive exploration of a particular plant without adequate rehabilitation efforts e.g. *Kayu Eben* (*Diospyros celebica*) in Sulawesi. Indirectly, for example [52], forest destruction due to air pollution or acid rain in industrialized countries such as the *Picea abies* plant in West Germany with damage of about 9% in 1982 accelerated to 51% in 1984.

From the results of in-depth interviews in the field, the factor of the occurrence of scarcity of rare plant species in the *Alas Kedaton* forest tourism are: 1) Environmental degradation factors. In this context it means that the present forests, inherited by the younger generation of the village today, are the remnants of the ancient forest, which now extend to 27 hectares. While the former, the extent of more than today. The age of their parents in the past, many converted the forest into agricultural areas. The current forest area, left to not be felled because in it there *Pura* as a holy place for Hindus to pray. So that the remaining forests are now believed to be a sacred place for Hindus in Bali; 2) Plants belonging to the rare category, seen from the way of reproduction, take place very slowly, so that the parents are very uninteresting to breed it; Thus its proliferation only takes place naturally, and its survival also takes place naturally; 3) Plants that are included in rare plants, have a high enough quality of wood, so many plants that live outside the forest of this tour, felled and used for building materials; 4) Rare plants considered as "sacred wood" by the people, often used for holy shrines (*Hindu* temples in Bali) or for religious ceremonies (*Hindu*), are not accompanied by breeding or breeding as materials Replacement of harvested plants; 5) The absence of an attempt to breed rare plants by forest managers and by surrounding communities. This effort is not done related to the increasingly difficult to find rare plants around their environment. Although the economic value is quite high, but because it is very rarely found in nature, the community turns to other timber, which is more practical, interesting and qualified to be used as a building material or as a reforestation material; 6) In the *Alas Kedaton* tourism forest, many rare plants are also dead, due to the disturbance of animals, especially long-tailed monkeys (*Macaca fascicularis*) and bats (*Pteropus vampyrus*), which are increasingly population. Herbs are often used as a place to play and many plants are "disturbed". Ecologically edible fruits are used as feed by some of the monkey populations and bats in the forest. Seeds that grow are often eaten or disturbed or broken so that the seeds of these plants to death.

Based on the average of effectiveness percentage of use of learning media to introduce of rare plants in *Alas Kedaton* tourism forest, Tabanan, Bali in terms of the overall component of CIPP evaluation model, it can be explained that in general the learning media can already function and good categorized because overall the average percentage of effectiveness level if viewed from all components obtained percentage of 88.20%. The results are reinforced with and proven from the average percentage of effectiveness level on the *context* component of 89.07% so that included in good category, the input component of 84.40% so that included in good category, on the process component of 90.13% so that

included in the category very good, And on product component equal to 88.80% so that included in good category.

Based on the comparison between the measurement results of effectiveness (shown in Table VI) with measurement standard of effectiveness (shown in Table V), so the learning media can be said to be effective on aspect-1, because the score of measurement results on aspect-1 was 89.60% having a higher score than the score of measurement standard amount of 88.00%. The learning media can be said to be effective on aspect-2, because the score of measurement results on aspect-2 was 92.00% having a higher score than the score of measurement standard amount of 90.00%. The learning media can be said to be effective on aspect-3, because the score of measurement results on aspect-2 was 85.60% having a higher score than the score of measurement standard amount of 85.00%. The learning media can be said to be effective on aspect-4, because the score of measurement results on aspect-2 was 83.20% having a higher score than the score of measurement standard amount of 80.00%. The learning media can be said to be effective on aspect-5, because the score of measurement results on aspect-5 was 89.60% having a higher score than the score of measurement standard amount of 88.00%. The learning media can be said to be effective on aspect-6, because the score of measurement results on aspect-6 was 81.60% having a higher score than the score of measurement standard amount of 80.00%. The learning media can be said to be effective on aspect-7, because the score of measurement results on aspect-7 was 82.40% having a higher score than the score of measurement standard amount of 80.00%. The learning media can be said to be effective on aspect-8, because the score of measurement results on aspect-8 was 88.80% having a higher score than the score of measurement standard amount of 85.00%. The learning media can be said to be effective on aspect-9, because the score of measurement results on aspect-9 was 92.00% having a higher score than the score of measurement standard amount of 90.00%. The learning media can be said to be effective on aspect-10, because the score of measurement results on aspect-10 was 89.60% having a higher score than the score of measurement standard amount of 85.00%. The learning media can be said to be effective on aspect-11, because the score of measurement results on aspect-11 was 80.80% having a higher score than the score of measurement standard amount of 80.00%. The learning media can be said to be effective on aspect-12, because the score of measurement results on aspect-12 was 88.00% having a higher score than the score of measurement standard amount of 85.00%. The learning media can be said to be effective on aspect-13, because the score of measurement results on aspect-13 was 91.20% having a higher score than the score of measurement standard amount of 88.00%. The learning media can be said to be effective on aspect-14, because the score of measurement results on aspect-14 was 87.20% having a higher score than the score of measurement standard amount of 85.00%. The learning media can be said to be effective on aspect-15, because the score of measurement results on aspect-15 was 86.40% having a higher score than the score of measurement standard amount of 85.00%. Obstacles found in this research is that the knowledge base is still limited from the results of explorative research on rare plants in the *Alas Kedaton* forest tourism only, whereas in

other forests there are actually other species that may have the same characteristics or even provide more complete information than obtained in the *Alas Kedaton* tourist forest.

IV. CONCLUSIONS

There are 48 species of plants that generally exist in the *Alas Kedaton* forest tourism. Of these, there are 42 (87.5%) plant species belonging to the rare category. Of the 42 species of rare plants present in the *Alas Kedaton* forest, there are 8 (19.04%) plant species belonging to the national rare category, 20 (47.62%) of rare plant species in Bali, 10 (23.81%) rare plant species in Tabanan District, and 4 (9.52%) species falling into the rare category at the Sub District level (especially Marga Sub-district).

The factors causing the scarcity of plant species present in the *Alas Kedaton* tourism forest are: 1) the degradation of the ancient environment, 2) the problem of reproduction of rare plants, 3) Human Intervention, 4) Disorders by animals, especially long-tailed monkeys (*Macaca fascicularis*) and bats (*Pteropus vampyrus*). Level of effectiveness of utilization of learning media to introduce of rare plants in *Alas Kedaton* tourism forest for students and general public is categorized in good category because in whole if evaluated from all component of evaluation CIPP model obtained the average effectiveness percentage of 88.20%.

Future works that can be done to overcome the constraints related to the knowledge base is finding out a source of knowledge based on explorative research in other locations and find sources in books or other related literature either from libraries or through the internet and also can develop applications with the use of data mining concept.

ACKNOWLEDGMENTS

The authors would like to extend their gratitude to all faculty members of Universitas Pendidikan Ganesha and IKIP PGRI Bali, who assisted in the completion of this research.

REFERENCES

- [1] Direktorat Jendral Planologi Kehutanan. (2010). *Data dan Kesatuan Pengelolaan Hutan 2010*. Jakarta: Kementerian Kehutanan Republik Indonesia.
- [2] Departemen Kehutanan. (2010). *Risalah Hutan Lindung di Wilayah KPH Bali Tengah dan Bali Timur Dinas Kehutanan Provinsi Dati 1 Bali*. Singaraja: Balai Inventarisasi dan Perpetaan Hutan.
- [3] Wijana, N., and Sumardika. I.N. (2004). *Penentuan Kualitas Air Danau Batur (Kajian dari Sisi ABC Envirotment*. Singaraja: Universitas Pendidikan Ganesha.
- [4] Wijana, N., and Sumardika. I.N. (2005). *Analisis Vegetasi Hutan Bukit Kangin Desa Adat Tenganan Pengringsingan, Kabupaten Karangasem*. Singaraja: Universitas Pendidikan Ganesha.
- [5] Wijana, N., Sutajaya, M. and Kariasa. N. (2006). *Analisis Kualitas Air, Aspek Kesehatan Masyarakat, Vegetasi Penyangga dan Upaya Pengelolaan oleh Masyarakat Seputar Danau Batur Kecamatan Kintamani Kabupaten Bangli*. Singaraja: Universitas Pendidikan Ganesha.
- [6] Wijana, N. (2008). "Keanekaragaman Spesies Tumbuhan, Manfaat dan Upaya Pelestariannya," *Jurnal Matematika dan Sains*, Vol. 5, No. 10, pp.17-34.
- [7] Wijana, N., and I.N. Sumardika, I.N. (2009). "Pelestarian Jenis-Jenis Tumbuhan Berguna Melalui Kearifan Lokal di Desa Adat Tenganan Pegringsingan, Kabupaten Karangasem, Bali," *Proceeding Konservasi Flora Indonesia Dalam Mengatasi Dampak Pemanasan Global*, pp. 724- 731.

- [8] Wijana, N., Swasta, I.B.J., and Sumardika, I.N. (2010). *Analisis ABC Environment Pada Ekosistem Tumpang Tindih (Overlap Ecosystem) dalam Kaitannya dengan Penurunan Kualitas Air, Eutrofikasi, dan Rencana Pemantauan Lingkungan (RPL) Danau Buyan Kecamatan Sukasada Kabupaten Buleleng*. Singaraja: Universitas Pendidikan Ganesha, 2010.
- [9] Wijana, N. (2012). *Analisis Dampak Lingkungan dan Upaya Pengelolaan Berbasis Ergologi Kawasan Wisata Lovina, Buleleng – Bali*. Singaraja: Universitas Pendidikan Ganesha.
- [10] Wijana, N. (2013). *Analisis Vegetasi Hutan Adat, Upaya Pengelolaan Berbasis Kearifan Lokal dan Pemberdayaan Masyarakat Melalui Pendekatan Ergologi di Desa Bali Aga Buleleng-Bali*. Singaraja: Universitas Pendidikan Ganesha.
- [11] Wijana, N. (2014). “Analisis Komposisi dan Keanekaragaman Spesies Tumbuhan di Hutan Desa Bali Aga Tigawasa, Buleleng-Bali,” *Jurnal Sains dan Humaniora Lemlit Undiksha*, Vol. 1, No. 1, pp. 55-65.
- [12] Wijana, N. (2015). *Analisis Dampak Lingkungan Terhadap Aktivitas Pembudidayaan Udang dengan Sistem Kurungan di Laut Lepas Desa Sangsit Kecamatan Sawan, Kabupaten Buleleng, Bali*. Singaraja: Universitas Pendidikan Ganesha.
- [13] Wijana, N. (2016). *Pengelolaan Lingkungan Hidup (Aspek Kearifan Lokal, Ergonomi, Ergologi, dan Regulasi)*. Singaraja: Undiksha Press.
- [14] Wijana, N. (2016). *Analisis Kualitas Lingkungan ditinjau dari aspek ABC Environment di Kawasan Wisata Toya Bungkah, Bangli*. Singaraja: Universitas Pendidikan Ganesha.
- [15] Wijana, N. (2016). *Penentuan Titik-titik Rawan Erosi Sepanjang Jalur Wisata Bedugul-Singaraja*. Singaraja: Universitas Pendidikan Ganesha.
- [16] Arrijani, Setiadi, D., Guhardja, E., and Qayim, I. (2006). “Analisis Vegetasi Hulu DAS Cianjur Taman Nasional Gunung Gede-Pangrango,” *Jurnal Biodiversitas*, Vol.7, No. 2, pp. 147-153.
- [17] Hartini, S. (2007). “Keragaman Flora dari Monumen Alam Kersik Luway Kalimantan Timur,” *Jurnal Biodiversitas*, Vol. 8, No. 1, pp. 67-72.
- [18] Purwaningsih. (2006). “Analisis Vegetasi Hutan pada Beberapa Ketinggian Tempat di Bukit Wawouwai, Pulau Wawonii Sulawesi Tenggara,” *Jurnal Biodiversitas*, Vol. 7, No. 1, pp. 49-53.
- [19] Purwaningsih, and Yusuf, R. (2008). “Analisis Vegetasi Hutan Pegunungan di Taman Nasional Gunung Ciremai, Majalengka, Jawa Barat,” *Jurnal Biologi Indonesia*, Vol. 4, No. 5, pp.385-399.
- [20] Junaedi, Indrawan, D., and Mutaqien, Z. (2010). “Diversity of Tree Communities in Mount Patuha Region, West Java”, *Biodiversitas*, Vol. 11, No. 2, pp. 75-81.
- [21] Onrizal, Kusmono, C., Saharjo, B. H., Handayani, I.P., and Koto, T. (2006). “Analisis Vegetasi Hutan Hujan Tropika Dataran Rendah Sekunder di Taman Nasional Danau Sentarum, Kalimantan Barat,” *Jurnal Biologi*, Vol. 4, No. 6, pp. 359-371.
- [22] Sujaya, I.M. (2007). Warga Kukuh Pantang Tebang Pohon di Alas Kedaton (access from <http://www.balisaja.com/2007/12/warga-kukuh-pantang-tebang-pohon-di.html>)
- [23] Divayana, D.G.H., and Sanjaya, D.B. (2017). “Mobile Phone-Based CIPP Evaluation Model in Evaluating the Use of Blended Learning at School in Bali,” *International Journal of Interactive Mobile Technologies*, Vol. 11, No. 4, pp.149-159.
- [24] Jampel, I.N., Lasmawan, I.W., Ardana, I.M., Ariawan, I.P.W., Sugiarta, I.M., & Divayana, D.G.H. (2017). “Evaluation of Learning Programs and Computer Certification at Course Institute in Bali Using CSE-UCLA Based on SAW Simulation Model,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 24, pp. 6934-6949.
- [25] Arnyana, I.B.P., Sadia, I.W., Suma, I.K., and Divayana, D.G.H. (2017). “Determination of Effectiveness of Evaluation Results on School Culture and Character of Junior High School Students Using Character Assessment Instruments With The Local Wisdom of Bali Based on Mobile Phone,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 20, pp. 5348-5359.
- [26] Divayana, D.G.H., Sanjaya, D.B., Marhaeni, A.A.I.N., and Sudirtha, I.G. (2017). “CIPP Evaluation Model Based on Mobile Phone in Evaluating The Use of Blended Learning Platforms at Vocational Schools in Bali,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No 9, pp. 1983-1995.
- [27] Divayana, D.G.H., Agung, A.A.G., Sappaile, B.I., Simatupang, W., Sastrawijaya, Y., Sundayana, I.M., and Sugiharni, G.A.D. (2017). “Utilization of Open Source Technology in Determining of Validity and Reliability of Evaluation Model Instruments Based on ANEKA Values in Order to Evaluate The Quality of Computer Learning,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 20, pp. 5517- 5534.
- [28] Divayana, D.G.H., Sappaile, B.I., Pujawan, I.G.N., Dibia, I.K., Artaningsih, L., Sundayana, I.M., & Sugiharni, G.A.D. (2017). “An Evaluation of Instructional Process of Expert System Course Program by Using Mobile Technology-Based CSE-UCLA Model,” *International Journal of Interactive Mobile Technologies*, Vol. 11, No. 6, pp. 18- 31.
- [29] Divayana, D.G.H., Marhaeni, A.A.I.N., Dantes, N., Arnyana, I.B.P., and W. Rahayu, (2017). “Evaluation of Blended Learning Process of Expert System Course Program by Using CSE-UCLA Model Based on Mobile Technology”, *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 13, 2017, pp. 3075-3086.
- [30] Ariawan, I.P.W, Sanjaya, D.B., and Divayana, D.G.H. (2016). “An Evaluation of the Implementation of Practice Teaching Program for Prospective Teachers at Ganesha University of Education Based on CIPP-Forward Chaining,” *International Journal of Advanced Research in Artificial Intelligence*, Vol. 5, No. 2, pp. 1-5.
- [31] Divayana, D.G.H., Ardana, I.M., and Ariawan, I.P.W. (2017). “Measurement of Effectiveness of a Lecturer in Transferring Algebra Knowledge Through of Multimedia Facilities by Using Certainty Factor-Formative-Summative Model,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No 9, pp. 1963-1973.
- [32] Divayana, D.G.H. (2015). “Evaluasi Program Penanggulangan HIV/AIDS Dengan Model CIPP Berbantuan Komputer,” *Konferensi Nasional Sistem & Informatika*, pp.442-446.
- [33] Divayana. D.G.H. (2016). *Evaluasi Program Perpustakaan Digital Berbasis Sistem Pakar pada Universitas Teknologi Indonesia*. Jakarta: Universitas Negeri Jakarta.
- [34] Divayana, D.G.H. (2015). “Penggunaan Model CSE-UCLA Dalam Mengevaluasi Kualitas Program Aplikasi Sistem Pakar,” *SNATIA*, pp.165-168.
- [35] Divayana, D.G.H. (2017). “Evaluasi Pemanfaatan E-Learning Menggunakan Model CSE-UCLA,” *Jurnal Cakrawala Pendidikan*, Vol. 36, No. 2, pp. 280-289.
- [36] Divayana, D.G.H. (2017). “Utilization of CSE-UCLA Model in Evaluating of Digital Library Program Based on Expert System at Universitas Teknologi Indonesia: A Model for Evaluating of Information Technology-Based Education Services,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 15, pp. 3585-3596.
- [37] Sanjaya, D.B., and Divayana, D.G.H. (2015). “An Expert System-Based Evaluation of Civics Education as a Means of Character Education Based on Local Culture in the Universities in Buleleng,” *International Journal of Advanced Research in Artificial Intelligence*, Vol. 4, No. 12, pp. 17-21.
- [38] Divayana, D.G.H., and Sugiharni, G.A.D. (2016). “Evaluasi Program Sertifikasi Komputer Pada Universitas Teknologi Indonesia Menggunakan Model CSE-UCLA,” *Jurnal Pendidikan Indonesia*, Vol. 5, No. 2, pp. 865-872.
- [39] Divayana, D.G.H., Adiarta, A., and Abadi, I.B.G.S. (2017). “Conceptual and Physical Design of Evaluation Program for Optimizing Digital Library Services at Computer College in Bali Based on CSE-UCLA Model Modification with Weighted Product,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 16, pp. 3767-3782.
- [40] Suandi, I.N., Putrayasa, I.B., and Divayana, D.G.H. (2017). “Compiling a Dictionary of Loan Words in Balinese: The Evaluation Result of Effectiveness Testing in The Field Aided by Mobile Technology,” *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 14, pp. 3186-3195.
- [41] Divayana, D.G.H., Adiarta, A., and Abadi, I.B.G.S. (2018). “Initial Draft of CSE-UCLA Evaluation Model Based on Weighted Product in Order to Optimize Digital Library Services in Computer College in Bali,” *IOP Conference Series: Materials Science and Engineering*, Vol. 296, 2018, pp. 12-17.

- [42] Sudiana, I.K., Rahayu, W., Santiyadnya, N., Parmithi, N.N, Mahendra, I.W.E., and Divayana, D.G.H. (2018). "Mapping Sports Tourism in Buleleng-Bali Using Goal-oriented Evaluation Model Based on SAW," *Journal of Theoretical and Applied Information Technology*, Vol. 96, No. 13, 2018, pp. 4157-4169.
- [43] Mahayukti, G.A., Dantes, N., Candiasa, I.M., Marhaeni, A.A.I.N., Gita, I.N., and Divayana, D.G.H. (2018). "Computer-based Portfolio Assessment to Enhance Students' Self-Regulated Learning", *Journal of Theoretical and Applied Information Technology*, Vol. 96, No. 8, 2018, pp. 2351- 2360.
- [44] Barbour, M.G., Burk, J.H., and Pitts, W.D. (1987). *Terrestrial Plant Ecology*. California: The Benjamin/Cummings Publishing Company Inc.
- [45] Subana, M., and Sudrajat. (2001). *Dasar-Dasar Penelitian Ilmiah*. Bandung: CV. Pustaka Pelajar.
- [46] Wijana, N. (2018). *Ensiklopedia Floristik Tumbuhan Langka*. Yogyakarta: Plantaxia.
- [47] Wijana, N., and Setiawan, I.G.A.N. (2017). *Pemetaan Pencaran dan Pola Sebaran Spesies Tumbuhan Langka Serta Upaya Pengelolaan Berbasis Kearifan Lokal Pada Hutan Wisata di Propinsi Bali (Laporan Hasil Penelitian)*. Singaraja: LPPM Universitas Pendidikan Ganesha.
- [48] Peraturan Pemerintah Republik Indonesia No. 7 Tahun 1999.
- [49] Indriyanto. (2008). *Ekologi Hutan Cetakan ke-2*. Jakarta: Bumi Aksara.
- [50] Sugita, W. (2015). *Dampak Perkembangan Pariwisata Terhadap Keberlanjutan Usahatani Rumput Laut di Desa Kutuh, Kuta Selatan Kabupaten Badung*. Denpasar: Universitas Udayana.
- [51] Sarna, K. (1993). *Inventarisasi dan Pelestarian Tanaman Langka di Bali dalam Usaha Menunjang Obyek Wisata dan Studi*. Singaraja: FKIP Unud.
- [52] Soemarwoto, O. (1991). *Indonesia Dalam Kancah Isu Lingkungan Global*. Jakarta: PT. Gramedia Pustaka Utama.

Energy Consumption Evaluation of AODV and AOMDV Routing Protocols in Mobile Ad-Hoc Networks

Fawaz Mahiuob Mohammed Mokbal
College of Computer Science
Beijing University of Technology
Beijing, China

Khalid Saeed
Department of Computer Science
Shaheed Benazir Bhutto University
Sheringal Dir Upper, Pakistan

Wang Dan
College of Computer Science
Beijing University of Technology
Beijing, China

Abstract—Mobile Ad-hoc Networks (MANETs) are mobile, multi-hop wireless networks that can be set up anytime, anywhere without the need of pre-existing infrastructure. Due to its dynamic topology the main challenge in such networks is to design dynamic routing protocols, which are efficient in terms of consumption of energy and producing less overhead. The main emphasis of this research is upon the prominent issues of MANETs such as energy efficiency and scalability along with some traditional performance metrics for performance evaluation. Two proactive routing protocols used in this research are single-path AODV versus multi-path AOMDV. Extensive simulation has been done in NS2 simulator, which includes ten scenarios. The simulation results revealed that the performance of AOMDV is more optimal as compared to AODV in terms of throughput, packet delivery fraction and end to end delay. However, in terms of consumption of energy and NRL the AODV protocol performed better as compared to AOMDV.

Keywords—MANETs; routing protocols; AODV; AOMDV; energy efficiency; routing performance

I. INTRODUCTION

In Mobile Ad-hoc Networks (MANETs) the mobile nodes can connect dynamically using a variety of wireless media without any centralized infrastructure [1]. There are many advantages of MANETs as compared to the traditional network such as ease of establishment of network, reduced infrastructure cost etc. In MANETs each mobile node not solely operates as a host but additionally works as a router and has the capability to perform routing [2]. The transmission range of the mobile nodes is limited due to which the nodes frequently join and leave the network and as a result, the network topology updates again and again [3]. The mobility of the nodes in MANETs can cause the links to break due to which the nodes recalculate routing information in order to establish the links. This process consumes power, processing time, memory and produces additional traffic [4]. The potential of the Ad-hoc networks is that it can be used in the situations where infrastructure is not available and technically not possible to deploy such as disaster and military operations. The situations can also include low power sensor networks [5].

In MANETs, routing is a very critical task that should be deal with very care. To send the data between the source node and the destination as well as to establish the connection, there

is a need for routing protocols. Due to dynamic and unexpected topology changes in MANETs, the design of an efficient routing protocol in terms of consumption of energy and producing less overhead is very important and it is a significant challenge for such type of networks. The routing protocols have been developed to deal with the challenges, such as security, energy and delay. However, there are shortcomings in some aspects and improvement in others. Furthermore, the cooperative routing algorithms that are associated to energy gathering are quite limited [6].

There are few studies related to energy consumption calculation of wireless network in ad-hoc mode such as the research conducted in [7] proposed an energy efficient MAC protocol having multichannel and provisioning of quality of service in MANETs. The research conducted in [8] proposed an energy efficient secure selection of MPR mechanism which considers both security metrics as well as energy metrics for the selection of MPR. More specifically, there is a lack of detailed evaluation of energy consumption of mobile ad-hoc network protocols.

We believe that energy-aware designing and analysis of known-protocols for the ad-hoc networking environment needs sensible data of the energy consumption behavior of actual wireless nodes. Additionally, it's vital to present this information in a manner that is helpful to protocol developers as well as to researchers. The main focus of this research are benchmarking performance against criteria of energy efficiency and scalability along with most traditional performance metrics for performance evaluation of two proactive routing protocols. The first one is a single path named AODV [9, 10], while the second one is multi-path called AOMDV [11, 12] respectively. This research work provides a paradigm for future studies of the development of dynamic routing protocols which are more efficient in terms of energy consumption and producing less overhead. All of which are considered to be prominent issues of MANETs. This research utilized the mendeley reference manager [13] for organizing this research, as well as for referencing.

A. MANETs Routing Protocols

Generally, MANETs routing protocols are often categorized into the subsequent three categories [14, 15]:

1) Proactive or Table-Driven routing protocols are based on the traditional link state and distance vector algorithms that are primarily meant for wired networks. These protocols maintained and periodically update their routing tables through interchanging the broadcast control messages.

2) Reactive or On-demand routing protocols are designed to have less overhead as compared to proactive routing protocols because the connection is only established when it is required by the source. This is typically done through a two-stage process known as route discovery.

3) In order to increase the overall scalability of routing Hybrid routing protocols were introduced which includes the features of both reactive and proactive routing protocols. In hybrid routing protocols, the network is comprised of various zones. The network route within each zone is kept up proactively and the routes between zones are resolved responsively.

B. Ad hoc On-Demand Distance Vector (AODV) Routing Protocol

AODV [9, 10] is proactive, single path, loop-free distance vector routing protocol. It is based on DSR's on-demand route discovery mechanism, with the idea of destination sequence numbers from DSDV, but it is different from DSDV by using hop-by-hop routing approach. AODV maintains routes only between nodes which need to communicate with each other. Each mobile node keeps a routing table which maintains information about next-hop of a path towards the destination node. In order to transport packets correctly towards the destination, the protocol uses two procedures: Route discovery of route between the source and the destination and route maintenance. It uses route request message (RREQ) and route reply message (RREP) for route discovery, and uses Route Error (RERR) for route maintenance. Moreover, Hello messages are used to preserve the connectivity between neighboring nodes.

C. Ad hoc On-Demand Multipath Distance Vector Routing Protocol

Depending upon the distance vector idea and utilizing hop-to-hop routing concept AOMDV discover routes on-demand utilizing a route discovery technique. The primary difference between AOMDV and AODV lies within the number of routes found in every route discovery [11]. The essence of the AOMDV protocol lies on guaranteeing that multiple paths discovered will be loop-free as well as disjoint, and in effectively finding such routes utilizing a flood-based route discovery [12]. Route update runs in AOMDV are executed locally at each node which has a key part in keeping up loop-freedom and disjoint attributes.

The rest of the paper is organized as follows: Section II includes the relevant research work done. Section III contains the research methodology adopted for carrying out this research. In Section IV results generated are discussed in detail. In Section V the research work done is concluded and at the end there are references.

II. LITERATURE REVIEW

Although energy consumption is agreed to be of importance within the design of ad-hoc networks routing protocols. However, most of studies regarding performance evaluation relied on traditional performance parameters such as throughput, end-to-end delay, PDF and NRL. Moreover, there is a great need to investigate the energy consumption of known protocols in MANETs for future researches.

The research that has been done in MANETs follows two trends, The first trend is the research work related to the design of efficient ad-hoc routing protocols aiming to achieve one or a combination of the targets such as increase in the packet delivery, minimizing energy consumption, and reducing the overheads in MANETs [16]-[24]. However, there are shortcomings in some aspects and improvement in others. The second, upon which the vast majority of research focuses, is performance evaluation based on traditional performance metrics [25]-[31].

The research conducted in [16] proposed a novel fault-tolerant routing approach utilizing a stochastic learning-based weak estimation procedure. The proposed scheme aims to make routing protocol successfully operate in adversarial environment. Authors in [17] have tried to reduce the waste of the limited battery power that occur in exchanging cluster maintenance messages by assigning critical node that has highest priority to be selected as a cluster head, as a results, limited battery power is preserved.

The research conducted in [18] proposed a dynamic energy efficiency algorithm which aims to extend the network lifespan, the proposed approach used two threshold ,yellow threshold that was used to obtain some sort of local load balancing via distributing the load equally among the neighboring cluster-heads, and a red thresholds that was used to prompt local re-clustering in the network. The result obtained in this research revealed that the proposed approach achieved better efficiency than those found in existing weight clustering approach.

The research conducted in [19] proposed a Bird Flight-Inspired Routing Protocol (BFIRP), the aim was to make highly scalable, dynamic, energy efficient, and position-based routing protocol. The proposal was based on three-dimensional (X, Y, Z) to determine the source and destination location. The outcomes demonstrate that the algorithm was highly scalable, and had low end-to-end delay compared to AODV as well as more efficient than AODV in terms of energy and throughput by 20% and 15% respectively.

The research conducted in [20] proposed learning automata based fault-tolerant routing algorithm which is able to perform routing in the existence of faulty nodes in MANETs. To achieve the optimize selection of paths, decrease the overhead in the network, and for learning about the faulty nodes existence in the network, they have utilized the theory of Learning Automata. The outcomes demonstrate that the packet delivery ratio increased and the overhead decreased as compared to the AODV protocol.

The research conducted in [21] proposed energy efficiency algorithm for a communication network in MANETs. The

proposal aims to optimize energy consumption through selecting the best path in terms of energy for transferring data after computing the energy required for each available path.

The research conducted in [22] proposed Ant-Colony Optimization (ACO) approach for selecting the optimal cluster heads. The aim was optimization of energy consumption as well as stability of the node. The probability function was used to compute the parameters like residual energy, energy drain rate and mobility factor. Node that has the highest value for the probability function will be selected as a cluster-head. The overall workload of communication is computed periodically. The cluster head is reset, if its value is high. The outcome shows that the approach has energy efficiency and clusters stability.

The research conducted in [23] attempted to decrease energy consumption and delay in MANETs. The proposed approach computed the important matrices such as Residual Energy, Node connectivity and Available Bandwidth for election of the cluster head efficiently. A conscious cluster routing algorithm was proposed by using constructed shortest path multicast tree that pick a cluster head as group leader and cluster members as leaf nodes. The most proposed approaches are extension of some of the current protocols which are either reactive protocols such as AODV and DSR or proactive protocols such as OLSR and DSDV.

The research conducted in [24] proposed the AOMDV-ER for improving of network lifetime and reduce routing overhead by using recoil off time technique based on their geographical location in order to reduce the number of transmissions. The outcomes show that the proposed scheme such as AOMDV-ER was able to save energy consumption up to 16%, and 12% reduction in routing overhead.

The second working trend are research on benchmarking and performance analysis of known network protocols, focused on traditional performance metrics such as PDF, throughput and End-to-End delay; or survey studies.

AODV and AOMDV in [25], [26] are compared with connections up to 50. They have concluded that AOMDV has more routing overhead and delay as compared to AODV, but it has better efficiency in packet drop and PDF.

The research conducted in [27] evaluated the performance of DSR, AODV and AOMDV routing protocol in MANETs by comparing the PDR, throughput, and end-to-end delay. They observed that in a network with increased number of nodes up to 20 nodes, PDF and throughput in AOMDV and DSR routing protocols are better as compared to AODV whereas the delay is less in AOMDV as compared to DSR and AODV.

The survey conducted in [28] reviewed typical reactive routing protocols and revealed the characteristics and trade-offs of AODV, AODV-UU, AOMDV, DSR and DYMO. They have concluded that each of the protocol in the conducted research performs well in some cases and has certain drawbacks in others scenarios.

The performance of AODV, AOMDV, DSR and DSDV were evaluated in [29] through comparing the PDR, packet loss ratio, and end-to-end delay performance matrices for wireless networks. They observed that the performance of AODV is best as compared to AOMDV, DSDV and DSR and therefore

the performance of DSR is best as compared to AODV, AOMDV and DSDV in TCP connection type as well as in CBR connection type.

The research conducted in [30] compared and analyzed the performance of AODV and AOMDV routing protocols in MANETs relying on the traditional performance metrics like throughput, end-to-end delay, PDF. They have observed that AOMDV performs well as compared to AODV in terms of PDF and throughput, however, AOMDV incurs a lot of delay in comparison to AODV.

The research conducted in [31] includes AODV, AOMDV, DSDV and DSR. They examined the effect of dynamic change in network topology on the performance based on traditional metrics such as PDR, end-to-end delay and NRL. They observed that AOMDV and DSDV are not suitable when the network topology updates again and again, while AODV and DSR are suitable in such scenario. DSR and DSDV performed better as compared to other protocols in terms of packet delivery ratio, end-to-end delay and NRL. AOMDV had less end-to-end delay but when the network topology changes more frequently, the PDF and NRL are worst as compared to other protocols.

III. RESEARCH METHODOLOGY

This research is based on evaluating the performance of AODV and AOMDV routing protocols in varied aspects, especially in energy consumption. To evaluate the performance, these protocols are simulated using NS-2 version 35 (The Network Simulator - ns-2, <https://www.isi.edu/nsnam/ns>) [32]. The simulation workflow is shown in Fig. 1.

A. Simulation Environment

NS2.35 is an object oriented simulator, which is built by combining the advantages of C++ with an OTcl languages. NS2 has full supports for multi-hop wireless ad-hoc environment integrated with physical, data link, and medium access control (MAC) layer model [33]. This research utilized these advantages of NS simulator to set and configure the environment for this research. The protocols have a send buffer of 64 packets to maintain the data packets start with route discovery phase, which are waiting to get the route that has not yet arrived. The mechanism that prevents unlimited buffering is to drop packets in buffer that took longer than 30 seconds. The interface queue that has a maximum size of 50 packets is used to maintain the routing layer packets that are sent until the MAC layer transmits them. The interface queue has two priorities for packets, each perform FIFO order mechanism. The higher primacy is given to routing packets as opposed to data packets [34].

The evaluations in this research depends on the simulation of 10, 20, 30, 40 and 50 wireless nodes for each protocol, moving randomly along a simulation area (800m x 800m) flat grid for 100 seconds simulation time. A square field grants nodes to move freely with a similar density. For the sake of a fair comparison between the two protocols, we have made the same environment and the same parameters for both protocols mentioned in Table I. Fig. 2 shows the simulation environment setup and configuration.

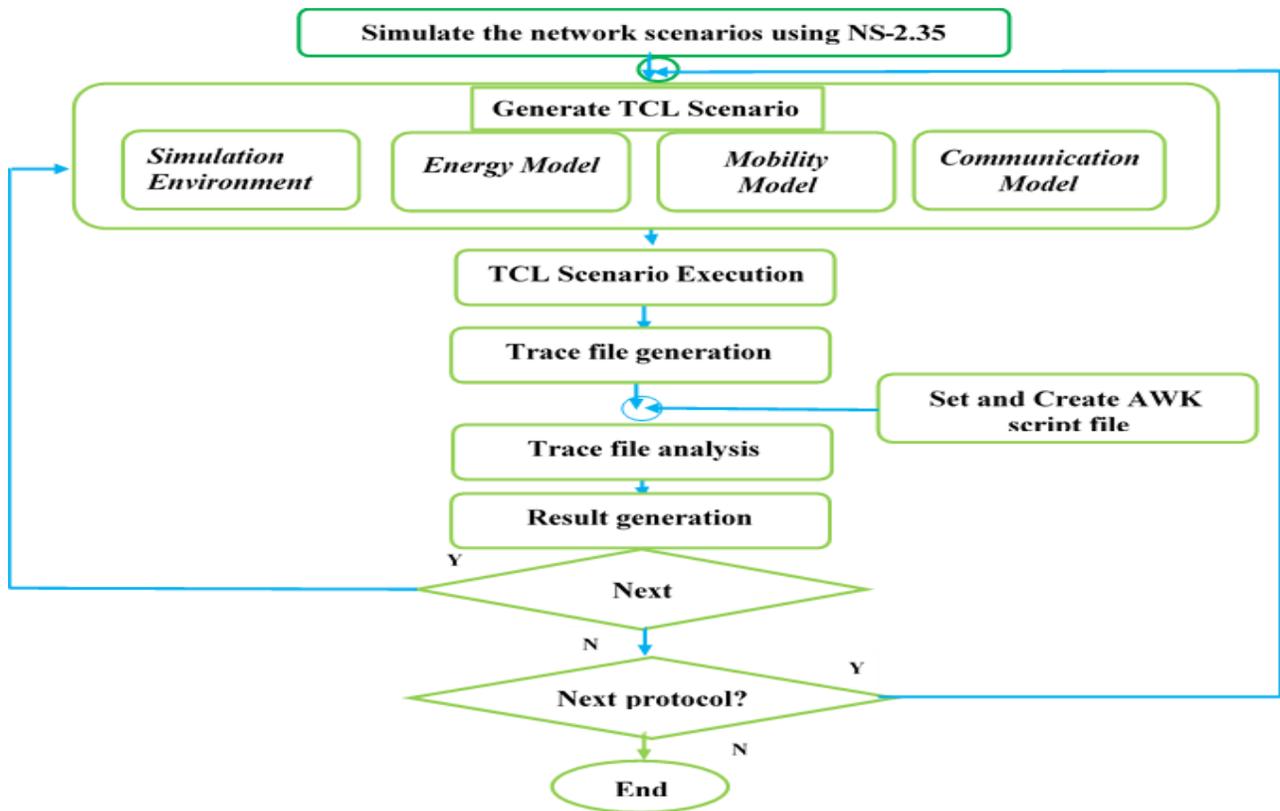


Fig. 1. Simulation Workflow.

```

1. set val(chan) Channel/WirelessChannel ; # channel type
2. set val(prop) Propagation/TwoRayGround ; # radio-propagation model
3. set val(ant) Antenna/OmniAntenna ; # Antenna type
4. set val(ll) LL ; # Link layer type
5. set val(ifq) Qucuc/DropTail/PriQucuc ; # Interface queuc type
6. set val(ifqlen) 50 ; # max packet in ifq
7. set val(netif) Phy/WirelessPhy ; # network interface type
8. set val(mac) Mac/802_11 ; # MAC type
9. set val(nn) 50 ; # number of mobilenodes
10. set val(rp) AOMDV ; # routing protocol
11. set val(x) 800 #Simulation area_X
12. set val(y) 800 #Simulation area_Y
13. set opt(energymodel) EnergyModel ; # Energy Model
14. set opt(initialenergy) 100 ; #initial energy in joules
15. set ns [new Simulator]
16. set f [open aomdv_50.tr w] #trace file
17. $ns trace-all $f
18. set namtrace [open aomdv_50.nam w] #nam file
19. $ns namtrace-all-wireless $namtrace $val(x) $val(y)
20. set topo [new Topography]
21. Stopo load_flatgrid 800 800
22. create-god $val(nn)
  
```

Fig. 2. Simulation Environment Configuration.

B. Energy Model

The parameters of energy model used in this research are mentioned in Table I and its detail is reflected in Fig. 3. The energy model is used to measure the power consumed in each scenario. The node consumes the available energy (initial energy) based on the following parameters: (1) Transmission

(Tx) (2) Reception (Rx) (3) Idle (4) Sleep (5) TransitionPower and (6) TransitionTime states. Transmission manner indicates the energy consumed (Watt) for transferring each packet, reception manner indicates the energy consumed (Watt) for receiving each packet, idle manner indicates the energy consumed (Watt) when the node is in idle mode, sleep manner indicates the energy consumed (Watt) when the node is in sleep mode, TransitionPower indicates the energy consumed (Watt) in case of transition from sleep to idle. TransitionTime indicates the time (second) which is used in case of transition from sleep to idle.

TABLE I. SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
Network Simulator	NS2.35	Transition Power	0.2 W
Type of Channel	Wireless Channel	Transition Time	0.005 S
Radio Propagation Model	Two Ray Ground	Routing Protocols	AODV, AOMDV
Type of Antenna	Omni Antenna	Mobility Model	Random Waypoint
Type of Interface queue	DropTail/PriQueue	Simulation Time	100 seconds
Max Packet in Ifqueue	50	Number of Scenarios	10 (5x2)
Type of Network Interface	Phy/WirelessPhy	Number of Nodes	10,20,30,40,50

Type of MAC layer	Mac/802.11	Transport Layer Protocol	UDP (User Datagram Protocol)
Simulation Area	800m x 800m	Traffic Model	CBR (Constant Bit Rate)
Initial Energy for Each Node	100 Joule	Packet Size	512 bytes
Transmission Power	2.0 W	Link Capacity	1.0 Mbps
Reception Power	1.0 W	Connection Rate	4 packets/sec
Idle Power	0.5 W	Number of Connections	1,2,3,4,5
Sleep Power	0.001 W	Node Speed	10m/s

```

1. $ns node-config -adhocRouting $val(rp) \
2. -llType $val(ll) \
3. -macType $val(mac) \
4. -ifqType $val(ifq) \
5. -ifqlen $val(ifqlen) \
6. -antType $val(ant) \
7. -propType $val(prop) \
8. -phyType $val(netif) \
9. -channel $chan_1 \
10. -topoInstance Stopo \
11. -agentTrace ON \
12. -routerTrace ON \
13. -macTrace ON \
14. -movementTrace ON \
15. -energyModel Sopt(energymodel) \
16.     -idlePower 0.5 \
17.     -rxPower 1.0 \
18.     -txPower 2.0 \
19.     -sleepPower 0.001 \
20.     -transitionPower 0.2 \
21.     -transitionTime 0.005 \
22.     -initialEnergy Sopt(initialenergy)

```

Fig. 3. Energy Model.

```

# dynamic rand destination procedure.
1. $ns at 0.5 "target"
2. proc target {} {
3.     global ns val n
4.     set time 1.0
5.     set now [$ns now]
6.     for {set i 0} {$i<$val(nn)} {incr i} {
7.         set xx [expr rand()*800]
8.         set yy [expr rand()*800]
9.         $ns at $now "$n($i) setdest $xx $yy 10.0"
10.    }
11.    $ns at [expr $now+$time] "target"
}

```

Fig. 4. Dynamic Mobility Function.

C. Mobility Model

In this research, the mobility model used is random waypoint, in which the mobile node move randomly and update their location, speed and acceleration change over time. It is simple and widely available model, thus, it is the most common mobility models to evaluate MANETs routing protocol [35]. In this research, node movement is done by

dynamic destination setting procedure as shown in Fig. 4. The scenario files utilized for each simulation are distinguished by same pause time which is 0.5s. All nodes start the simulation by remaining stationary for the pause time period. At the end of pause time period, the node randomly select destination in the simulation area, moving in space at a uniform speed of 10m/s for the entire period of the simulation.

D. Communication Model

This research used the traffic pattern to be constant bit rate (CBR) source over the User Datagram Protocol (UDP) at transport layer. The origin and target pairs are spread randomly across the network. Packets size 512 bytes is used, while the number of CBR packets generated vary based on the connection rate, Hence, for all scenarios in these simulations, we choose to fix connection rate at 4 packets/sec. Five different communication patterns corresponding to 1, 2, 3, 4 and 5 connections for 10, 20, 30, 40, 50 nodes respectively were considered. The communication pattern of 5 connections is shown in Fig. 5.

```

#procedur CBR agent
1. proc attach-CBR-traffic { node sink size interval } {
#Get an instance of the simulator
2.     set ns [Simulator instance]
#Create a CBR agent and attach it to the node
3.     set cbr [new Agent/CBR]
4.     $ns attach-agent $node $cbr
5.     $cbr set packetSize_ $size
6.     $cbr set interval_ $interval
7.     $cbr set rate_ 1.0Mb
#Attach CBR source to sink;
8.     $ns connect $cbr $sink
9.     return $cbr
10. }
11. set cbr0 [attach-CBR-traffic $n(0) $sink1 512 0.25]
12. set cbr1 [attach-CBR-traffic $n(7) $sink2 512 0.25]
13. set cbr2 [attach-CBR-traffic $n(20) $sink3 512 0.25]
14. set cbr3 [attach-CBR-traffic $n(30) $sink4 512 0.25]
15. set cbr4 [attach-CBR-traffic $n(40) $sink5 512 0.25]
16. $ns at 1.0 "$cbr0 start"
17. $ns at 1.0 "$cbr1 start"
18. $ns at 1.0 "$cbr2 start"
19. $ns at 1.0 "$cbr3 start"
20. $ns at 1.0 "$cbr4 start"
21. $ns at 100.0 "finish"
22. puts "Start of simulation..."
23. $ns run

```

Fig. 5. Communication Pattern.

E. Performance Metrics Used in Simulation

In order to evaluate the performance of AODV and AOMDV, we considered the eight most commonly used quantitative indicators to judge the performance of the protocols: (1) Total Energy consumed by all nodes (TE); (2) Average Consumed Energy (ACE); (3) Average Residual Energy (ARE); (4) Packet Delivery Fraction (PDF); (5) Throughput Rate [kbps]; (6) End-to-End delay (E2ED); (7) Routing Load and (8) Normalized Routing Load.

1) Total Energy consumed by all nodes (TE): Energy consumption is computed as follows:

The time needed for transmitting a data packet is

$$Time = 8 \times (Psize/BW)$$

Therefore, we have:

$$E_{ti} = P_{ti} \times Time$$

$$E_{ri} = P_{ri} \times Time.$$

$$E_{ldi} = P_{ldi} \times Time$$

The transition power mode with transition time(t) is as follows: $E_{tpi} = P_{tpi} \times \text{Time}(t)$

Where E_{ti} indicates the amount of energy consumed by a node i in the transmission power mode, E_{ri} indicate the amount of energy consumed by a node i in the Reception Power mode, E_{ldi} indicates the amount of energy consumed by a node i in the Idle mode, E_{si} indicates the amount of energy consumed by a node i in the sleeping mode, and E_{tpi} indicates the amount of energy consumed by a node i in the TransitionPower mode with TransitionTime (t) which is used for transition from sleep to idle. The total energy consumed by a node i is calculated as:

$$\text{TotalE } i = E_{ti} + E_{ri} + E_{ldi} + E_{ldi} + E_{si} + E_{t} \quad (1)$$

The Total Energy consumed (TE) by all nodes (N) is:

$$TE = \sum_{i=0}^N \text{TotalE } i \quad (2)$$

2) *Average Consumed Energy (ACE)*: It refers to the ratio of total energy consumed by each nodes (TE) to the number of nodes (N).

$$ACE = \frac{TE}{N} \quad (3)$$

3) *Average Residual Energy (ARE)*: It refers to the ratio of total initial energy of all nodes (IE) – total energy consumed by all nodes (TE) divided by number of nodes N.

$$ARE = \frac{\sum_{i=0}^N IE - \sum_{i=0}^N TE}{N} \quad (4)$$

4) *Packet Delivery Fraction (PDF)*: It indicates the ratio of correctly received packets to all sent packets in a period. It is an appraisal indicator of the reliability of transmission in Ad-Hoc network. The smaller value the packet delivery shows the worst performance.

$$PDF = \frac{\sum_{i=0}^N Pri}{\sum_{i=0}^N Psi} \quad (5)$$

N is the total number of nodes, Pri is the number of packets received by node i , Psi is the number of packets sent by node i .

5) *Throughput Rate [kbps] (TR)*: It points to the total received packets' size successfully reached at target per unit time.

$$TR [kbps] = \frac{\sum \text{received size}}{\text{StopT} - \text{StartT}} \times \frac{8}{1000} \quad (6)$$

6) *End-to-End Delay (E2ED)*: The time taken by the data packets to be arrived at destination sent by the source is known as Average End-to-End Delay. The Average End-to-End delay value refers to the time used for all potential delays results in buffering procedure, interface queuing, the retransmission procedure executed at MAC and propagation times. The lower the delay time, the better the efficiency.

$$\text{Average E2ED} = \frac{1}{NP} \sum_{i=0}^{NP} (Rt(i) - St(i)) \quad (7)$$

Where NP refers to total number of the packets received successfully, Rti points to the time when the packet i is received, Sti points to the time when the packet was sent.

7) *Routing Load*: The total routing packets transmitted including the packets which are forwarded at network layer are known as Routing Load.

$$\text{Routing Laod} = CPSn + CPFn \quad (8)$$

where, CPSn points to the number of routing control packets generated to be sent, CPFn points to the number of routing control packets to be forwarded, CPSn and COFn at network layer.

8) *Normalized Routing Load (NRL)*: Normalized routing load is the average number of routing control packets transmitted at network layer per data packets received by destination at the application layer. It refers to the congestion status of the network. The higher routing load increases the probability of network congestion.

$$NRL = \frac{\text{Routing Laod}}{DPn} \quad (9)$$

Where DPn refers the total number of the data packets received.

IV. RESULTS AND DISCUSSION

This section includes the details discussion about the results generated during simulation. In this research for analyzing the trace file for each scenario the AWK scripting language [36], [37] is used. Fig. 6 illustrates the energy tracking function of nodes which uses the trace files generated through simulation as input and store the output in the matrix. While Fig. 7 illustrates compute function of energy that is consumed by nodes, which uses the output of tracking function as input and compute consumed energy for each node as output.

Tables II, III, IV, V and VI show the results obtained regarding energy consumption by each node in the various scenarios separately for both protocols AODV and AOMDV.

```

1. BEGIN {
2.   initialenergy = 100
3.   energy_left[size] = initialenergy
4.   maxenergy=0
5.   totalenergy=0
6.   total=0
7.   n=50
8.   n1=0
9.   nodeid=0
10.  i=0;
11.  }
12.  {
13.    event = $1
14.    time = $2
15.    if (event=="N"){
16.      node_id = $5
17.      energy=$7
18.      for(i=0;i<n;i++) {
19.        if(i == node_id) {
20.          finalenergy[i]=energy;
21.        }
22.      }
23.    }
24.    if(event == "N") {
25.      for(i=0;i<n;i++) {
26.        if(i == node_id) {
27.          energy_left[i] = energy_left[i] - (energy_left[i] - energy);
28.        }
29.      }
30.    }
31.  }

```

Fig. 6. Energy Tracking Function.

```

1. BEGIN {
2.     # Compute consumed energy for each node
3.     for (i in finalenergy) {
4.         consumenergy [i]=initialenergy -finalenergy [i]
5.         totalenergy +=consumenergy [i]
6.         if(maxenergy <consumenergy [i]){
7.             maxenergy=consumenergy [i]
8.             nodeid=i
9.         }
10.    }
11.    #compute average energy
12.    averagenergy=totalenergy/n
13.    #output
14.    for (i=0; i<n; i++) {
15.        print("node",i, consumenergy [i])
16.    }
17.    print("average",averagenergy)
18.    print("total energy",totalenergy)
19.    for(i=0;i<n;i++) {
20.        total=total+energy_left [i];
21.        if(energy_left [i]!=0)
22.            n1++;
23.    }
24.    n1--;
25.    printf("Average residual energy :%.6f\n", total/n);
26. }

```

Fig. 7. Compute Consumed Energy Function.

TABLE II. ENERGY CONSUMPTION BY EACH NODE IN 10 NODES SCENARIO

Node No.	AODV	AOMDV	Node No.	AODV	AOMDV
0	55.2753	55.8414	5	55.5218	56.0896
1	53.4916	54.0586	6	53.2687	53.8353
2	53.2687	53.8403	7	53.2143	53.7748
3	55.5221	56.0913	8	53.2687	53.8362
4	53.2687	53.7633	9	53.2687	53.8386

TABLE III. ENERGY CONSUMPTION BY EACH NODE IN 20 NODES SCENARIO

Node No.	AODV	AOMDV	Node No.	AODV	AOMDV
0	59.7815	59.6455	10	57.8163	57.7453
1	58.0402	57.9685	11	57.8163	57.7374
2	58.0594	57.9863	12	57.8163	57.7341
3	57.8163	57.7386	13	57.6738	57.6145
4	60.0735	57.7366	14	60.0701	57.7366
5	57.8163	57.7341	15	57.8163	59.9957
6	60.0697	57.7332	16	60.0704	57.7378
7	59.8023	59.7733	17	57.8163	57.7357
8	57.8163	59.9987	18	60.0613	59.9865
9	57.8163	60.0051	19	57.8163	57.7378

TABLE IV. ENERGY CONSUMPTION BY EACH NODE IN 30 NODES SCENARIO

Node No.	AODV	AOMDV	Node No.	AODV	AOMDV
0	62.0289	62.3113	15	62.3425	60.4472
1	62.4216	60.6681	16	60.0786	62.7112
2	60.3226	60.6946	17	60.0786	60.4405
3	60.0786	62.71	18	62.3304	60.4111
4	60.2165	60.4458	19	60.0786	60.442
5	60.0786	60.4421	20	62.1021	62.4772
6	60.078	60.3897	21	60.3209	60.7035
7	62.0824	62.3957	22	60.0786	60.4405
8	60.0786	60.4496	23	60.0786	60.4428
9	62.2012	60.4561	24	60.0786	60.4429
10	60.0786	60.4412	25	60.0786	60.4491
11	62.3355	60.4604	26	62.3341	62.7078
12	60.0786	60.4413	27	60.0786	60.4471
13	59.975	62.686	28	59.9649	60.2529
14	60.0786	62.7056	29	60.0265	60.3115

TABLE V. ENERGY CONSUMPTION BY EACH NODE IN 40 NODES SCENARIO

Node No.	AODV	AOMDV	Node No.	AODV	AOMDV
0	68.8375	68.2596	20	68.9956	68.4826
1	71.4571	66.6842	21	67.2109	66.6952
2	69.331	66.7028	22	66.9651	66.4381
3	69.0919	66.452	23	66.9685	66.4482
4	69.3759	66.4544	24	71.4909	66.4494
5	67.1073	66.4397	25	66.9685	66.4499
6	66.9337	68.6657	26	69.242	66.4486
7	68.8342	68.4365	27	66.9679	66.4493
8	66.9685	68.7176	28	69.085	66.1894
9	66.9685	68.7181	29	66.862	66.3122
10	66.9685	68.7128	30	68.8081	68.3569
11	66.9685	70.9735	31	67.0836	66.5254
12	69.2285	66.4527	32	66.9685	66.4462
13	66.7829	68.6642	33	66.9658	66.4346
14	66.9685	66.4478	34	66.9685	66.4486
15	66.9685	66.4478	35	66.9685	66.4487
16	69.2305	68.717	36	66.9563	66.4145
17	66.9679	66.4428	37	66.6084	66.2002
18	66.9229	66.328	38	66.9018	66.3646
19	67.1089	66.4519	39	66.9685	68.7203

TABLE VI. ENERGY CONSUMPTION BY EACH NODE IN 50 NODES SCENARIO

Node No.	AODV	AOMDV	Node No.	AODV	AOMDV
0	66.7606	73.3017	25	64.8803	73.8995
1	65.1066	71.875	26	64.8799	71.6341
2	67.2366	71.8762	27	67.1378	73.8461
3	64.973	71.6389	28	64.6579	71.2808
4	67.1531	71.6319	29	64.7657	71.3901
5	64.967	71.5672	30	66.7479	73.4913
6	64.8442	71.5385	31	65.0551	71.5351
7	66.8665	73.613	32	64.883	71.6359
8	64.8819	71.6327	33	64.8614	71.6213
9	67.1568	71.6373	34	64.8771	71.6133
10	67.1546	72.5123	35	64.8786	71.6206
11	64.882	71.6337	36	64.7843	73.8235
12	64.9353	71.6114	37	64.534	73.4686
13	64.7378	71.486	38	64.8162	71.5277
14	64.8811	76.7877	39	64.9392	73.899
15	67.149	71.6372	40	66.9181	73.6779
16	66.9909	73.9153	41	64.9904	71.3731
17	64.8782	71.5975	42	64.7865	71.5662
18	64.8756	73.8302	43	64.6201	70.7853
19	64.8773	71.6173	44	64.6883	71.4151
20	66.9138	76.5577	45	64.5232	70.5279
21	67.4006	72.4824	46	64.5299	71.1699
22	64.9339	73.8115	47	64.308	69.802
23	64.8771	71.6168	48	64.3463	70.7955
24	64.878	71.62	49	63.9482	68.8861

Table VII and VIII show the evaluation results obtained for the AODV in different scenarios used in this research. Table IX and X shows the evaluation results obtained for the AOMDV in different scenarios.

TABLE VII. ENERGY CONSUMPTION EVALUATION OF AODV

N	PS	PR	PD	TE	ACE	ARE
10	396	396	0	539.368	53.9368	46.0632
20	792	792	0	1171.87	58.5933	41.4067
30	1188	1187	1	1822.18	60.7394	39.2606
40	1584	1582	2	2710.98	67.7744	32.2256
50	1980	1976	4	3268.67	65.3734	34.6266

N: NO. OF NODES, PS: PACKET SENT, PR: PACKET RECEIVED, PD: PACKET DROPPED, TE: TOTAL ENERGY CONSUMED BY ALL NODES (JOULES), ACE: AVERAGE ENERGY CONSUMED BY EACH NODE, ARE: AVERAGE RESIDUAL ENERGY FOR EACH NODE.

TABLE VIII. PERFORMANCE EVALUATION OF AODV

N	PDF %	TR [bps]	TR kbps	E2ED (s)	RL	NRL
10	100	202752	16.42	0.017984	12	0.030
20	100	405504	32.84	0.0322659	45	0.057
30	99.9158	607744	49.21	0.0412604	159	0.134
40	99.8737	809984	65.55	0.140721	527	0.333
50	99.798	1011712	81.90	0.403008	664	0.336

N: NO. OF NODES, PDF: PACKET DELIVERY FRACTION, TR: THROUGHPUT RATE, E2ED: AVERAGE END-TO-END DELAY, RL: ROUTING LOAD, NRL: NORMALIZED ROUTING LOAD

TABLE IX. ENERGY CONSUMPTION EVALUATION OF AOMDV

N	PS	PR	PD	TE	ACE	ARE
10	396	396	0	544.969	54.4969	45.503058
20	792	792	0	1168.08	58.4041	41.595934
30	1188	1188	0	1830.82	61.0275	38.972501
40	1584	1583	1	2685.89	67.1473	32.852694
50	1980	1977	3	3609.32	72.1863	27.813691

N: NO. OF NODES, PS: PACKET SENT, PR: PACKET RECEIVED, PD: PACKET DROPPED, TE: TOTAL ENERGY CONSUMED BY ALL NODES (JOULES), ACE: AVERAGE ENERGY CONSUMED BY EACH NODE, ARE: AVERAGE RESIDUAL ENERGY FOR EACH NODE.

TABLE X. PERFORMANCE EVALUATION OF AOMDV

N	PDF%	TR [bps]	TR kbps	E2ED (s)	RL	NRL
10	100	202752	16.42	0.01812	1005	2.538
20	100	405504	32.84	0.02989	2027	2.559
30	100	608256	49.25	0.03975	3110	2.618
40	99.9369	810496	65.61	0.05884	4155	2.625
50	99.8485	101222	81.92	0.07327	5219	2.64

N: No. of Nodes, PDF: Packet Delivery Fraction, TR: Throughput Rate, E2ED: Average End-To-End Delay, RL: Routing Load, NRL: Normalized Routing Load

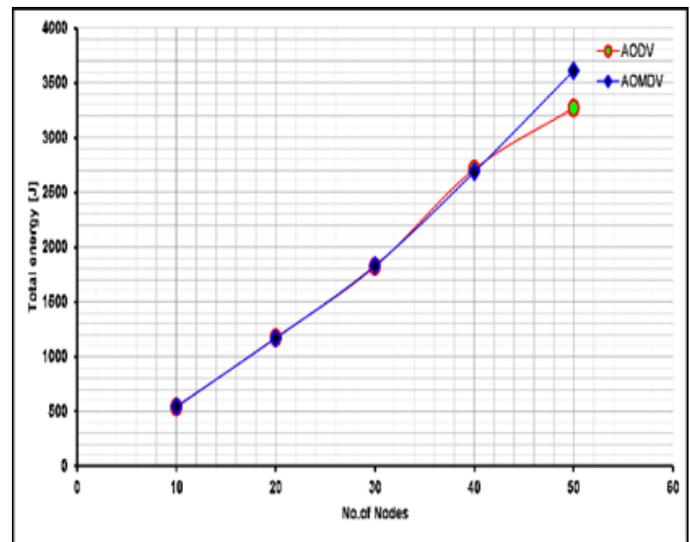


Fig. 8. Total Energy Consumed By All Nodes (Joules) Vs Number of Nodes.

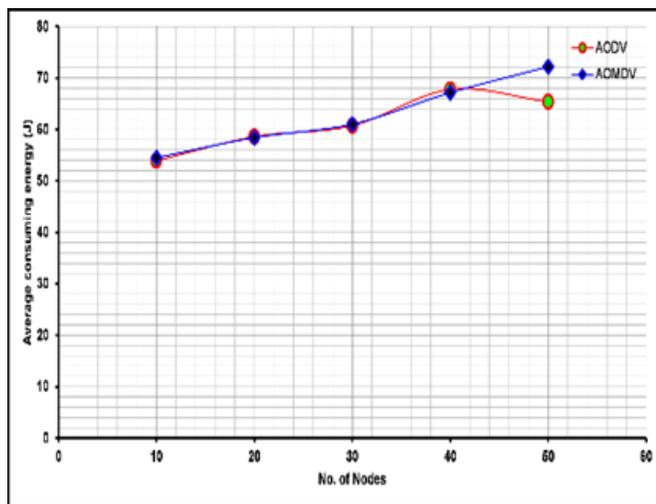


Fig. 9. Average Energy Consumed by each Nodes Vs Number of Node.

Fig. 8 shows the total energy consumed by all nodes of each scenario. The outcomes demonstrated that the AODV consumed less energy as compared to AOMDV, the possible reason behind this is that AODV is single-path protocol and found single path to destination due to which it consumed less energy. The term average energy consumed reflects the percentage of energy consumed by each node. Fig. 9 shows this result, which shows more energy consumed by the AOMDV when the number of nodes increased.

Fig. 10 shows the percentage of residual energy or battery life for each node in different scenarios, by using the equation number (4); it is clear from Fig. 10 that the AODV has more residual energy as compared to AOMDV. PDF indicates the percentage of packets that arrived at the destination successfully. Fig. 11 shows the PDF of AODV and AOMDV in the first two scenarios (at 10, 20 nodes with 1 and 2 connections) are almost same. However, with the increase in the number of nodes and CBR connections (at 30, 40, 50 nodes with 3, 4, 5 connections) AOMDV showed better results as compared to AODV.

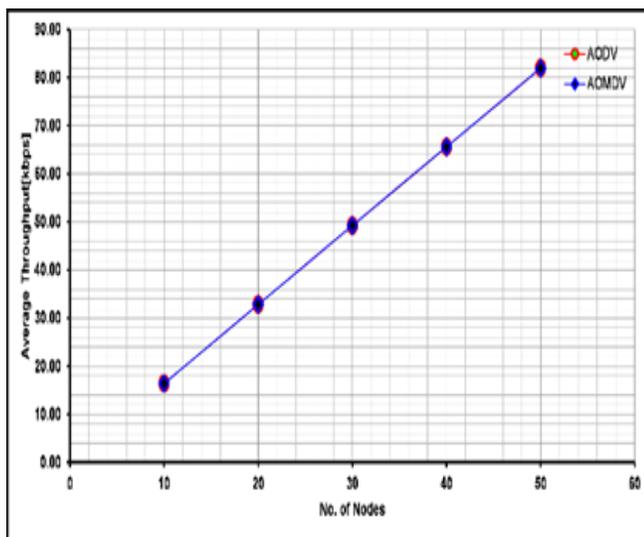


Fig. 10. Average Residual Energy Vs Number of Nodes.

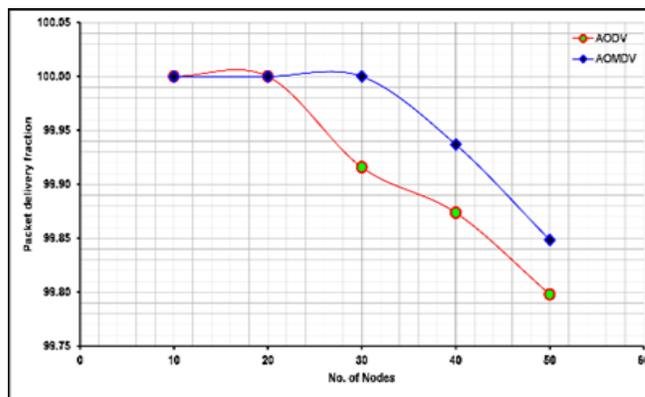


Fig. 11. Packet Delivery Fraction Vs Number of Nodes.

Fig. 12 shows the network throughput rate of AODV and AOMDV versus the number of nodes. Based on the results, AOMDV produced better throughput rate. In other words, when the number of nodes increased the AOMDV throughput increased and when the number of nodes decreased the AOMDV throughput decreased. AODV is a single-path routing protocol whose average end-to-end delay is higher as compared to multi-path protocols. Fig. 13 clearly shows the higher delay of AODV as the number of nodes and the number of connections increases, and in case of AOMDV it reduced. This is the nature of the AOMDV protocol, which works to find alternate paths when the basic path is lost without having to rediscover the path, and therefore does not require extra time.

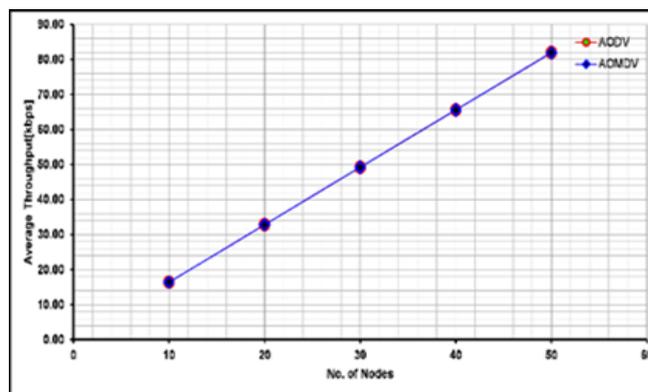


Fig. 12. Throughput Rate [kbps] Vs Number of Nodes.

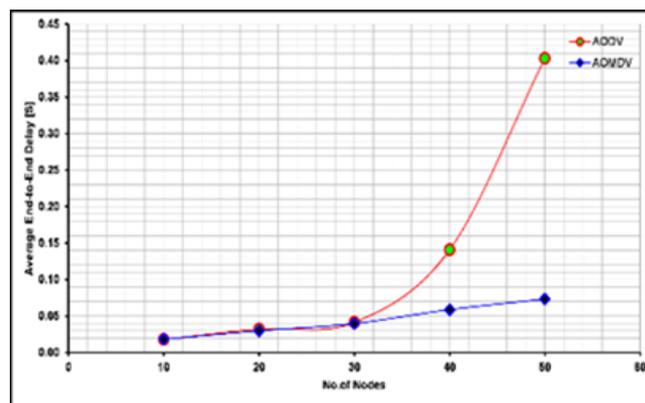


Fig. 13. Average End-to-End Delay Vs Number of Nodes.

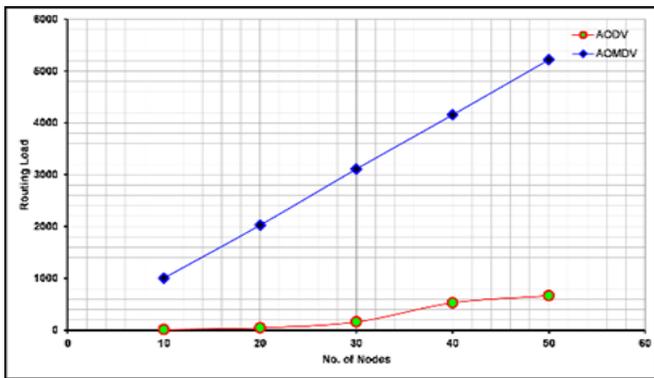


Fig. 14. Routing Load Vs Number of Nodes.

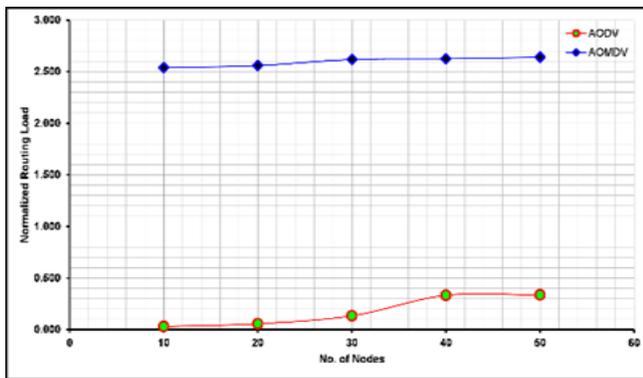


Fig. 15. Normalized Routing Load Vs Number of Nodes.

Routing load of AODV and AOMDV protocol is shown in Fig. 14. The simulation outcomes demonstrated that the AODV protocol produced less routing overhead because it is a single-path protocol. NRL indicates the number of routing packets transmitted including the forwarded packets per data packets delivered at application layer to the destination. Fig. 15 shows the simulation results of the NRL of AODV and AOMDV at different number of nodes, and in various CBR connections. It has been observed that AOMDV has higher NRL. The reason is that routing overhead is higher in AOMDV because the nature of the protocol is multi-path, where the routing packets seeking to find many alternate routes are retained and are used in case of loss of connection of the main path in order to reduce end-to-end delay and increases packet delivery rate.

V. CONCLUSION

In this research performance evaluation of two routing protocols AODV and AOMDV in MANETs has been done. There is a lack of detailed evaluation of energy consumption of mobile ad-hoc network protocols. Furthermore, there is a great need to investigate the energy consumption of known-protocols in MANETs for future research studies. The vast majority of studies concentrated on performance parameters based on traditional performance metrics. This research provides a paradigm for future studies for the development of dynamic routing protocols, which are more efficient and effective in terms of energy consumption and producing less overhead.

Extensive simulation has been done in NS2 simulator, which includes ten scenarios, five for each protocol; vary in density of nodes and traffic. It has been concluded in this

research that the performance of AOMDV is more optimal as compared to AODV in terms of packet delivery fraction, throughput and end-to-end delay. However, in terms of energy consumption and normalized routing load, AODV is more optimal as compared to AOMDV. It is also concluded that AOMDV is more suitable when the network is stable; however, its performance is reduced when the network topology changes frequently. Furthermore, there is a trade-off in AOMDV routing protocol between energy consumption on the one hand and PDF efficiency and throughput on the other hand.

REFERENCES

- [1] E. Alotaibi and B. Mukherjee, "A survey on routing algorithms for wireless Ad-Hoc and mesh networks", *Computer. Networks*, vol. 56, no. 2, pp. 940–965, 2012. doi: 10.1016/j.comnet.2011.10.011.
- [2] M. Zhang and P. H. J. Chong, "Performance Comparison of Flat and Cluster-Based Hierarchical Ad Hoc Routing with Entity and Group Mobility", in *Proc. of IEEE Conf. on Wireless Communications and Networking*, pp. 1–6, April 5, 2009. doi: 10.1109/WCNC.2009.4917894.
- [3] Y. Khamayseh, O. M. Darwish, and S. A. Wedian, "MA-AODV: Mobility Aware Routing Protocols for Mobile Ad Hoc Networks", in *Proc. of 4th IEEE Conf. on Systems and Networks Communications*, pp. 25–29, 2009. doi: 10.1109/ICSNC.2009.80.
- [4] W. Wang and C. Amza, "Motion-based routing for opportunistic ad-hoc networks", in *Proc. of 14th ACM Conf. on Modeling, analysis and simulation of wireless and mobile systems- MSWiM '11*, pp. 169–178, October 31, 2011. doi: 10.1145/2068897.2068928.
- [5] B. Malarkodi, P. Gopal, and B. Venkataramani, "Performance Evaluation of Adhoc Networks with Different Multicast Routing Protocols and Mobility Models", in *Proc. of IEEE Conf. on Advances in Recent Technologies in Communication and Computing*, pp. 81–84, October 27, 2009. doi: 10.1109/ARTCom.2009.29.
- [6] A. Sarkar and T. Senthil Murugan, "Routing protocols for wireless sensor networks: What the literature says?", *Alexandria Engineering Journal*. Faculty of Engineering, vol. 55, no. 4, pp. 3173–3183, December, 2016. doi: 10.1016/j.aej.2016.08.003.
- [7] S. M. Kamruzzaman and Md. Abdul Hamid, "An Energy Efficient Multichannel MAC Protocol for QoS Provisioning in MANETs," *KSII Transactions on Internet and Information Systems*, vol. 5, no. 4, pp. 684–702, 2011. DOI: 10.3837/tiis.2011.04.004
- [8] Anjali Anand, Rinkle Rani and Himanshu Aggarwal, "Energy Efficient and Secure Multipoint Relay Selection in Mobile Ad hoc Networks," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 4, pp. 1571–1589, 2016. DOI: 10.3837/tiis.2016.04.006
- [9] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing", *Proc. 2nd IEEE. Workshop on Mobile Computing Systems and Applications*, pp. 90–100, 1999. doi: 10.1109/MCSA.1999.749281.
- [10] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc on-demand distance vector (AODV) routing", *Internet Society, No. RFC 3561*, pp. 1–37, 2003.
- [11] M. K. Marina and S. R. Das, "On-demand multipath distance vector routing in ad hoc networks", in *Proc. Of 9th IEEE Conf. on Network Protocols(ICNP)*, Comput. Soc, pp. 14–23, November 11, 2001. doi: 10.1109/ICNP.2001.992756.
- [12] M. K. Marina and S. R. Das, "Ad hoc on-demand multipath distance vector routing", *Wireless Communications and Mobile Computing*, vol. 6, no. 7, pp. 969–988, November, 2006. doi: 10.1002/wcm.432.
- [13] The Mendeley Support Team, Free reference manager and PDF organizer, Available at: <http://www.mendeley.com/>.
- [14] M. Abolhasan, T. Wysocki, and E. Dutkiewicz, "A review of routing protocols for mobile ad hoc networks", *Ad Hoc Networks*, vol. 2, no. 1, pp. 1–22, 2004. doi: 10.1016/S1570-8705(03)00043-X.
- [15] N. H. Saeed, M. F. Abbod, and H. S. Al-Raweshidy, "MANET routing protocols taxonomy", in *Proc. of IEEE Conf. on Future Communication Networks*, pp. 123–128, April 2, 2012. doi: 10.1109/ICFCN.2012.6206854.

- [16] B. J. Oommen and S. Misra, "A Fault-Tolerant Routing Algorithm for Mobile Ad Hoc Networks Using a Stochastic Learning-Based Weak Estimation Procedure", in *Proc. of IEEE Conf. on Wireless and Mobile Computing, Networking and Communication (WiMob'2006)*, pp. 31–37, June 19, 2006. doi: 10.1109/WIMOB.2006.1696374.
- [17] C.-C. Tseng and K.-C. Chen, "A Clustering Algorithm to Produce Power-Efficient Architecture for (N, B)-Connected Ad Hoc Networks", in *Proc. of IEEE Conf. on Communications*, pp. 3497–3502, June 24, 2007. doi: 10.1109/ICC.2007.578.
- [18] H. Safa, O. Mirza, and H. Artail, "A Dynamic Energy Efficient Clustering Algorithm for MANETs", in *Proc. of IEEE Conf. on Wireless and Mobile Computing, Networking and Communications*, pp. 51–56, October 12, 2008. doi: 10.1109/WiMob.2008.67.
- [19] S. Misra and G. Rajesh, "Bird Flight-Inspired Routing Protocol for Mobile Ad Hoc Networks", *ACM Transactions on Autonomous and Adaptive Systems*, vol. 6, no. 4:25, pp. 1–37, October, 2011. doi: 10.1145/2019591.2019594.
- [20] S. Misra, P. V. Krishna, A. Bhiwal, A. S. Chawla, B. E. Wolfinger, and C. Lee, "A learning automata-based fault-tolerant routing algorithm for mobile ad hoc networks", *The Journal of Supercomputing*, vol. 62, no. 1, pp. 4–23, October, 2012. doi: 10.1007/s11227-011-0639-8.
- [21] A. Choukri, A. Habbani, and M. El Koutbi, "An energy efficient clustering algorithm for MANETs", in *Proc. of IEEE Conf. on Multimedia Computing and Systems (ICMCS)*, pp. 819–824, 2014. doi: 10.1109/ICMCS.2014.6911232.
- [22] J. John and R. Pushpalakshmi, "A reliable optimized clustering in MANET using Ant Colony algorithm", in *Proc. of IEEE Conf. on Communication and Signal Processing*, pp. 051–055, April 3, 2014. doi: 10.1109/ICCSP.2014.6949797.
- [23] S. B. Kulkarni and B. N. Yuvaraju, "Node connectivity, Energy and Bandwidth Aware Clustering Routing Algorithm for real-time traffic multicasting in MANET", in *Proc. of IEEE Conf. on Advance Computing Conference (IACC)*, pp. 760–763, June 12, 2015. doi: 10.1109/IADCC.2015.7154809.
- [24] R. Sahu and N. Chaudhari, "Energy Reduction Multipath Routing Protocol for MANET Using Recoil Technique", *Electronics*, vol. 7, no. 5, p. 56, April, 2018. doi: 10.3390/electronics7050056.
- [25] S. Biradar and K. Majumder, "Performance Evaluation and Comparison of AODV and AOMDV", *International Journal on Computer Science and Engineering*, vol. 2, no. 2, pp. 373–377, 2010. Available at <http://www.enggjournals.com/ijcse/doc/IJCSE10-02-02-47.pdf>.
- [26] J. Jacob and V. Seethalakshmi, "Performance analysis and enhancement of routing protocol in MANET", *International Journal of Modern Engineering*, vol. 2, no. 2, pp. 323–328, Mach, 2012. Available at: http://www.ijmer.com/papers/vol2_issue2/BA22323328.pdf.
- [27] T. K. Araghi, M. Zamani, and A. B. A. Mnaf, "Performance Analysis in Reactive Routing Protocols in Wireless Mobile Ad Hoc Networks Using DSR, AODV and AOMDV", in *Proc. of IEEE Conf. on Informatics and Creative Multimedia*, pp. 81–84, 2013. doi: 10.1109/ICICM.2013.62.
- [28] D. N. Patel, S. B. Patel, H. R. Kothadiya, P. D. Jethwa, and R. H. Jhaveri, "A survey of reactive routing protocols in MANET", in *Proc. of IEEE Conf. on Information Communication and Embedded Systems (ICICES2014)*, pp. 1–6, February 27, 2014. doi: 10.1109/ICICES.2014.7033833.
- [29] B. Paul, K. A. Bhuiyan, K. Fatema, and P. P. Das, "Analysis of AOMDV, AODV, DSR, and DSDV Routing Protocols for Wireless Sensor Network", in *Proc. of IEEE Conf. on Computational Intelligence and Communication Networks (ICICES)*, pp. 364–369, November 14, 2014. doi: 10.1109/CICN.2014.88.
- [30] B. Rekha and D. V. Ashoka, "Performance analysis of AODV and AOMDV routing protocols on scalability for MANETs", In *Emerging Research in Electronics, Computer Science and Technology*, pp. 173–181. Springer, New Delhi, 2014. https://doi.org/10.1007/978-81-322-1157-0_19.
- [31] D. Lei, T. Wang, and J. Li, "Performance Analysis and Comparison of Routing Protocols in Mobile Ad Hoc Network", in *Proc. of 5th IEEE Conf. on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, pp. 1533–1536, September 18, 2015. doi: 10.1109/IMCCC.2015.325.
- [32] Network Simulator NS2 and Network Animator NAM, Available at: <http://www.isi.edu/nsnam> (Accessed 12 July 2017)
- [33] T. Issariyakul and E. Hossain, *Introduction to Network Simulator NS2*, Springer Science and Business Media, 2nd Edition, Springer, New York, 2011. doi: 10.1007/978-0-387-71760-9.
- [34] K. Fall and K. Varadhan, *The ns Manual* (formerly ns Notes and Documentation). *The VINT project*, 47, 2005. Available at: <http://www.isi.edu/nsnam/ns/ns-documentation.html>.
- [35] J. Yoon, M. Liu, and B. Noble, "Random waypoint considered harmful", in *IEEE INFOCOM 2003, Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428)*, vol. 2, pp. 1312–1321, March 30, 2003. doi: 10.1109/INFCOM.2003.1208967.
- [36] D.B. Close, A.D. Robbins, P.H. Rubin and R. Stallman, *The GAWK manual*, 0.15 Edition, Free Software Foundation, Inc, Cambridge, MA, April, 1993.
- [37] A. D. Robbins, *AWK language programming*. 1.0 Edition, Free Software Foundation, Inc, Boston, UAS, 1996.

Piezoelectric based Biosignal Transmission using XBee

Mohammed Jalil¹, Mohamed Al Hamadi², Abdulla Saleh³, Omar Al Zaabi⁴, Soha Ahmed⁵, Walid Shakhathreh⁶,
Mahmoud Al Ahmad⁷

Electrical Engineering Department
United Arab Emirate University
AlAin, UAE

Abstract—This paper is showcasing the development of an innovative healthcare solution that will allow patient to be monitored remotely. The system utilizes a piezoelectric sheet sensor and XBee wireless communication protocol to collect and transmit heart beat pressure signal from human subject neck to a receiving node. Then, using signal processing techniques a set of important vital parameters such as heart rate, and blood pressure are extracted from the received signal. Those extracted parameters are needed to assess the human subject health continuously and timely. The architecture of our developed system, which enables wireless transmission of the raw acquired physiological signal, has three advantages over existing systems. First, it increases user's mobility because we employed XBee wireless communication protocol for signal transmission. Second, it increases the system usability since the user has to carry a single unit for signal acquisition while preprocessing is performed remotely. Third, it gives us more flexibility in acquiring various vital parameters with great accuracy since processing is done remotely with powerful computers.

Keywords—Piezoelectric; XBee; medical sensors; vital signs; remote health monitoring

I. INTRODUCTION

According to the World Health Organization (WHO) 2015 reports, the most common health complaint is cardiovascular disease for both Emiratis and expatriates [1]. Furthermore, the Health Authority Abu Dhabi (HAAD) latest statistics disclosed that the leading cause for expat women death is cancer, followed by cardiovascular disease while Emirati women leading cause for death is heart disease [1]. Thus, UAE health authorities are working hard to keep up with the growing number of population, the increasing burden of chronic diseases, the rising number of aging people and the expanding medical tourism in the region [1]. Reports showed that in 2013 alone, UAE healthcare expenses reached \$16.8 billion [2]. A study performed by the Emirates Cardiac Society surveyed more than 4,000 people reported important findings. They found that nearly nine out of ten people in UAE are at risk of cardiovascular disease and one out of three of them are ignorant of this matter [1]. Cardiovascular disease are the leading cause for death worldwide — taking 17.3 million lives yearly — and UAE is not excluded [1]. Treatment for cardiovascular disease in UAE currently account for 36% of the total healthcare expenditure [3]. The above mentioned

problems have been fueling the rapid and increased interest in wearable mobile sensors and wireless sensing networks for healthcare applications. It is expected that those two technologies could reduce healthcare expenditure and disease prevalence by facilitating continues health monitoring and early disease detection. In this paper, the design and development of wireless health monitoring system is presented. The system utilizes piezoelectric sheet sensor to acquire physiological signal and employs XBee wireless communication protocols to send the acquired raw signal to a receiving node for processing and vital parameter extraction.

The remainder of this paper is organized as follows: Section II gives an overview of existing wireless health monitoring systems; Section III describes the experimental setup and various wireless communication protocols considered; and Section IV describes the data processing, the parameters extraction algorithms and the results. We then present our conclusion in Sections V.

II. STATE OF THE ART

The employment of wearable, cheap, unobtrusive, noninvasive, and wireless sensors in healthcare applications has attracted research and industries attention equally nowadays. The result of this devoted attention was huge number of applications and various technologies and products integration. In this section, a summary of the developments made in wearable, wireless, and medical monitoring systems will be presented. Wireless Health Monitoring System (WHMS) usually consists of three main parts: Physiological Signal Acquisition Module (PSAM), Signal Processing Module (SPM), and Remote Monitoring Module (RMM). Each and every part of these modules consists of submodules. Fig. 1 illustrates the main and submodules for wireless health monitoring systems (WHMS). The first part of any WHMS is usually the physiological signal acquisition module (PSAM). This module consists of two submodules which are biosensors and wire or wireless transmission unit. The second module is the signal processing module (SPM). This module comprises three submodules: signal wire/less receiving unit; memory, central processing unit (CPU) and wire/less transmission unit. The third module is the remote monitoring module (RMM). This module consists of wire/less receiving unit, database and a reports generating mechanism.

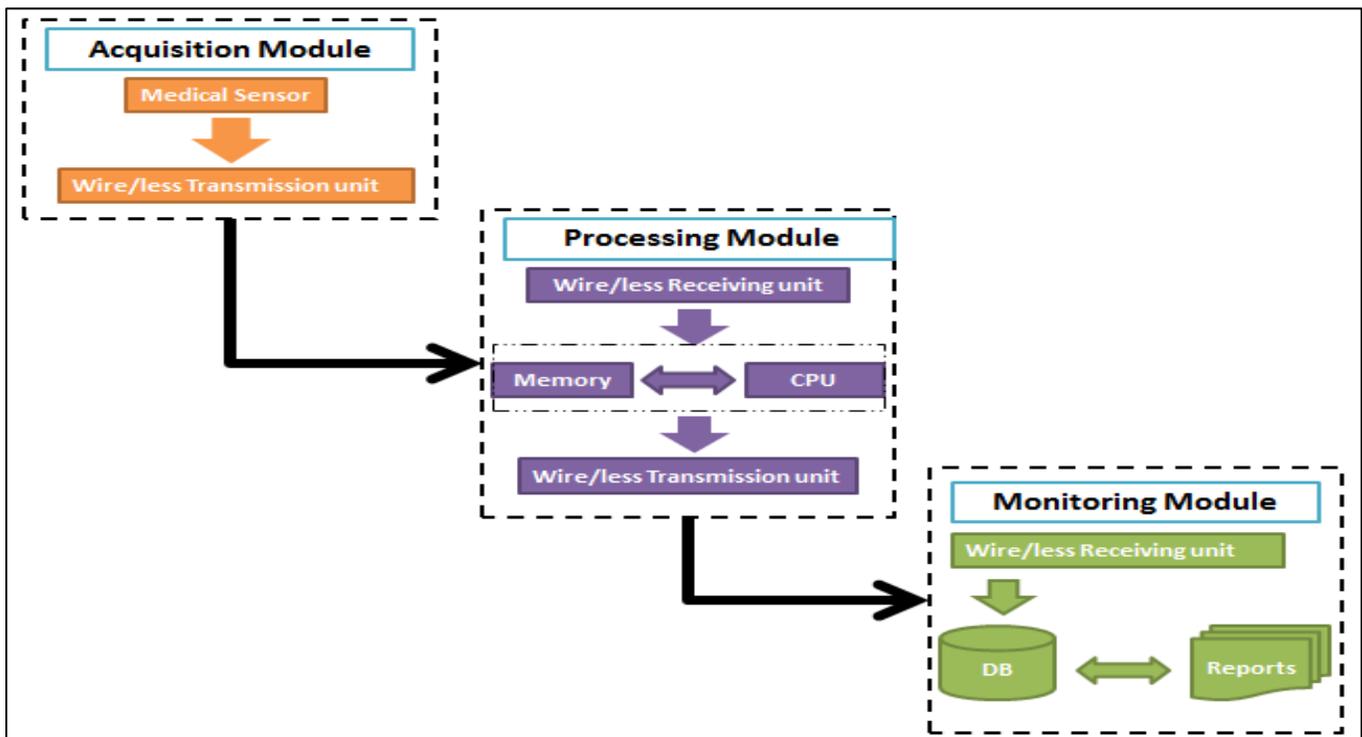


Fig. 1. Wireless Health Monitoring System (WHMS) general architecture.

In some systems, the data acquisition module sends the acquired data to the signal processing module (SPM) using wires like AMON System [4]. AMON combines the data acquisition module and the signal processing module in one component called wrist monitoring device. AMON uses wires to connect the two modules which is usually the case when the data acquisition module and the signal processing module are integrated in one component. Then, the raw signals and the extracted medical values can be sent to the remote monitoring module wirelessly using GSM technology. Another system that connect the PSAM and SPM via wires is the one developed by Sung-Nien and Jen-Chieh [5]. Their system employs two sensors to monitor patient health, namely, 1-lead ECG and respiration sensor. The acquired signals are sent to the SPM using wires and after processing the signal is sent to a local monitoring unit (placed at the patient room) via Bluetooth technology and from there to a remote monitoring unit via Wi-Fi network. Furthermore, Yuan-Hsiang et al. developed a WHMS for patient monitoring during transportation [6]. The system connects PSAM and SPM via RS232 wire connection. The system can be used by ambulance staff to monitor the patient and to inform remote medical staff (in the hospital) about the patient current medical status which will ensure that correct medical measurement is taken timely. The advantage of this system is that it utilizes existing PDA devices technology

to process the medical sensors data and to transmit the processed data wirelessly. Smart Vest system is a unique WHMS [7]. The system takes advantage of a washable shirt, which exploits an array of biosensors. The PSAM in Smart Vest utilizes ECG electrodes, temperature (Temp) sensor and photoplethysmography (PPG) sensor to acquire physiological signals. The SPM in Smart Vest is separated into two parts: the first part is connected to the PSAM via wires. This part performs several tasks for the raw physiological signal acquired such as filtering, amplification and digitization. The second part of the SPM is integrated with the RMM. The second part of SPM is responsible for deriving blood pressure (BP) by analyzing ECG and PPG waveform and extracting heart rate (HR) from ECG waveform. What is unique about this system is that it connects the two SPM parts wirelessly. A summary of the reviewed WHMS is presented in Table I. From the table, it can be seen that several vital signs has been monitored using those systems such as: blood pressure (BP), heart rate (HR), arterial oxygen saturation (SpO₂), temperature (Temp), respiration rate (RR), Galvanic Skin Response (GSR), etc. In addition, several wireless communication protocols were utilized such as the Global System for Mobile Communications (GSM), Bluetooth, WiFi, Wireless Local Area Network (WLAN), and Radio Frequency (RF).

TABLE I. SUMMARY OF WIRELESS HEALTH MONITORING SYSTEMS (WHMS)

The System	Sensors	Vital signs	Wireless transmission technology	PSAM and SPM
AMON System [4]	One-lead ECG, pulse oximeter, Blood Pressure, Acceleration Sensor, Temperature Sensor	BP, SpO2, HR, Temp (Optional: Glucose level and respiratory flow)	GSM	Combined in one unit and connected via wires
Wireless patient monitoring system [5]	one-lead ECG, respiration sensor	HR, RR	Bluetooth, WiFi network	Separate but connected via wires
Patient Transport system [6]	three-lead ECG, dual-wavelength photoplethysmographic (PPG) sensor	HR, SpO2	WLAN	Separate but connected via wires (RS232)
Smart Vest [7]	ECG electrodes, Photoplethysmogram (PPG) sensor, Thermistor	BP, Temp, HR, GSR	RF (Radio Frequency)	Separate connected via wires and wirelessly

Table II shows various research studies that were performed to measure heart rate or blood pressure or both. It summarizes the various methods used to measure important vital signs. As can be seen from the table, most research studies performed employed non-invasive methods because they are more convenient for use and require less drastic measures. Only one research study used invasive BP measurement [8]. Their argument is that this invasive method is very accurate and convenient for patients who require continuous monitoring and has critical condition. Furthermore, most of the studies focus on fabricating new transducers to measure vital signs [9]-[11]. Only three studies considered the effect of sending vital signs parameters wirelessly [12]-[15]. Nevertheless, those three studies performed the processing of the acquired vital sign signal in the source point and sent the interrupted vital sign data wirelessly. In other words, the signal processing was performed at the source location. From one hand, this made the acquired signal less susceptible to motion artifacts and noise but on the other hand, signal analysis was performed using modest processor with limited processing power and small memory.

Compared to the above-mentioned systems, our system has several unique features. As far as we know we are the only WHMS which connects the PSAM and SPM wirelessly via XBee. In other words, unlike the existing systems patients don't have to carry SPM around since the signal processing is done remotely. Our PSAM acquires the raw physiological signals and sends it using XBee wireless communication technology to the SPM. There are three advantages of this design approach namely: mobility, usability and flexibility.

Our system frees the patient from wearing or carrying around the SPM along with PSAM which is the case with the existing WHMS. As a result, the patient mobility will increase because the PSAM usually is light and the acquired signal is sent wirelessly. Second, sending the raw signals wirelessly to a remote SPM will give us more flexibility in signal processing. In other words, the raw signal can be analyzed with powerful processing units which will ensure the extraction of various vital sign information with great accuracy. While in the above-mentioned systems, signal preprocessing was performed via a

small-sized processing unit with modest memory and processing power. The processing unit had to be small since it is connected to the PSAM via wires otherwise the user will have to carry a big medical device. On the other hand, our system design architecture will increase its usability, since the patient has to carry a single unit which is the PSAM and hence he/she will be motivated to wear it continuously.

III. EXPERIMENTAL SETUP

In this section, the experiment design and setup will be described thoroughly. Fig. 2 illustrates the experimental setup for the proposed remote health monitoring system. Fig. 2(a) depicts a human subject placing piezoelectric sheet sensor on top of his carotid artery to sense the pressure pulse in this major artery. In Fig. 2(b) the transmitter module, the receiver module and a simple RC-filter are illustrated. The two modules are responsible of making a wireless connection using XBee protocol between the PSAM and the SPM.

The signal at the receiver node (XBee module) are filtered by a simple RC-filter to extract the analogue signals from the received Pulse Width Modulated (PWM) signal and with some digital signal processing, a set of vital parameters are extracted. The transmission module and the receiver module we used were XBee pro S1. Also, in the lab during the experiment we utilized the oscilloscope to display the acquired raw signal in the receiving point. In this setup, the stress signal sensed by the piezoelectric sheet was sampled at a sampling rate of 100Hz and converted into binary data and then assembled in frames and transmitted to the receiver module where they are filtered, processed and displayed on the oscilloscope. For our setup, we employed XBee communication protocol for many reasons. First, XBee protocol is a wireless communication standard for low data transmission rate and long distance. Thus, it is suitable for sensors and devices that do not require high data rate but needs long battery life, minimal user intervention and long distance. Second, XBee is convenient for different kind of applications such as medical, home/office automation and military applications. In addition, XBee networks may be implemented with several different and flexible network structures [18]-[21].

TABLE II. SUMMARY OF WIRELESS HEALTH MONITORING SYSTEMS (WHMS)

Paper	Procedure		Focus	Vital sign	Transducer		Transmission	Location
	Invasive	Non-invasive			Type	Material		
[9]		x	fabrication	BP	piezoelectric		NA	wrist
[10]		x	fabrication	BP	piezoelectric	EMF plastic	NA	upper arm
[11]		x	fabrication	BP	piezoelectric	ceramic bimorph beam	NA	wrist
[15]		x	fabrication	HR, RR	piezoelectric	aluminum nitride	NA	in bed
[12]		x	fabrication and Transmission	BP	piezoresistive		transmitting module employing a surface acoustic wave	upper arm
[8]	x		fabrication	SPO, BP	piezoelectric	cellular polypropylene	NA	directly at an arterial vessel
[16]		x	fabrication	HR	piezoelectric	PVDF polymer	NA	wrist
[17]		x	fabrication	Arterial Pulse Analyzers	piezoelectric	ceramic plate sensors	wired to a tablet	wrist
[18]		x	fabrication	BP	piezoelectric	zirconate titanate	NA	
[13]		x	Transmission	HR and BP	Commercial		XBee	upper arm
[14]		x	Transmission	BP	photoelectric plethysmography (PPG)		XBee	finger tip
[19]		x	Exploration	HR and BP	piezoelectric	sheet	NA	chest
[20]		x	fabrication	HR	piezoelectric	sheet	RF	ear

Table III depicts four wireless communication standards that are frequently used in WHMS. As can be seen in the table, XBee has low data rate in comparison to the other three wireless communication standards but it has a very long battery life, considerably long range and huge network structure can be built using XBee because the maximum number of nodes that can be accommodated in one network equal 65000. Thus, XBee is one of the most used wireless communication standards in ehealth applications [22]. For all the above-

mentioned reasons, we chose XBee wireless communication for our WHMS.

IV. RESULTS AND DISCUSSION

There are always tradeoffs between usability and reliability. The system architecture we exploited although improved patient mobility and increased system usability but it introduced noise to the acquired signal. The employment of XBee wireless communication protocol introduced a DC offset to the acquired signal.

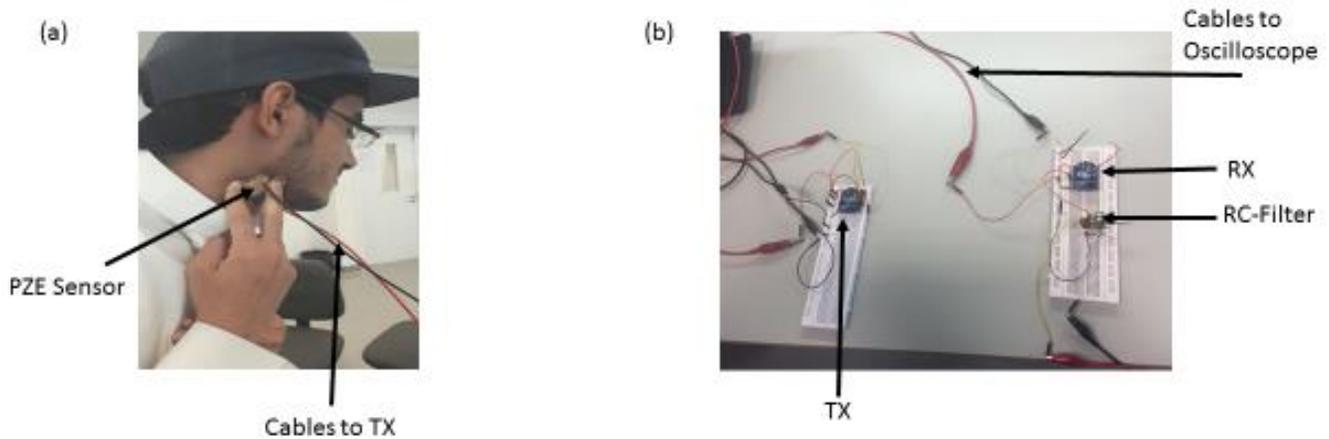


Fig. 2. Experiment setup (a) Piezoelectric sensor anchoring at the neck; (b) TX and RX XBee Modules.

TABLE III. WIRELESS COMMUNICATION STANDARDS

	IEEE Standard	Range (meter)	Maximum data rate (kbps)	Battery life (days)	Maximum Number of nodes
XBee	802.15.4	1-75+	20-250	100-1000+	65000
Bluetooth	802.15.1	1-10+	720	1-7	8
Wi-Fi	802.11b	1-100	11000+	1-5	32
GPRS/GSM	1XRTT/CDMA	1000+	64-128	1-7	1

At the SPM, a number of steps were taken to preprocess the acquired signal. First, a moving average filter with cutoff frequency of 20Hz was used to remove the white noise. Second, the signal was analyzed to detect the type of the DC offset that was introduced because of transmission using XBee wireless communication protocol. The detected DC offset trend was found to be nonlinear. Third, a low order polynomial fitting technique was used to remove the nonlinear trend from the signal.

After preprocessing, the data was analyzed and important parameters were extracted. The parameters extracted were the maximum and minimum absolute values and the time interval between peaks.

To calculate real-time heartbeat rate, we needed to find the average period of a measured cycle which was found to equal 0.824 second. This means using the resulting piezoelectric voltage signal the estimated heart rate per minute equal 72 beat per minute. To calculate the blood pressure and heart rate values we used the method explained in Saadat et al. work [19]. The method state that the piezoelectric sensor output voltage is directly proportional to the exerted pressure on the piezoelectric material. To calculate the pressure one must know

the equivalent turn ratio for the piezoelectric sensor [23]. The equivalent turn ratio for the piezoelectric material can easily be calculated from the information listed in the sensor data sheet [24]. Fig. 3 shows the resulting pressure signal extracted from the piezoelectric voltage signal.

The extraction of blood pressure using piezoelectric sensor will facilitate the continuous monitoring of heart rate and blood pressure.

V. CONCLUSION

This study is an attempt to develop an innovative healthcare solution that will allow patient to be monitored remotely. In particular two vital signs will be monitored namely heart rate and blood pressure. To this end, a piezoelectric sensor film was used to measure the pressure variance resulting from a heartbeat at the neck. This location has two advantages: good physiological signal to noise ratio (SNR) because carotid arteries are the major blood vessels that deliver blood to the brain, they are big and thus the pressure pulse waveform obtained from them will be very clear. Furthermore, a sensor placed in the neck is less prone to motion artifact.

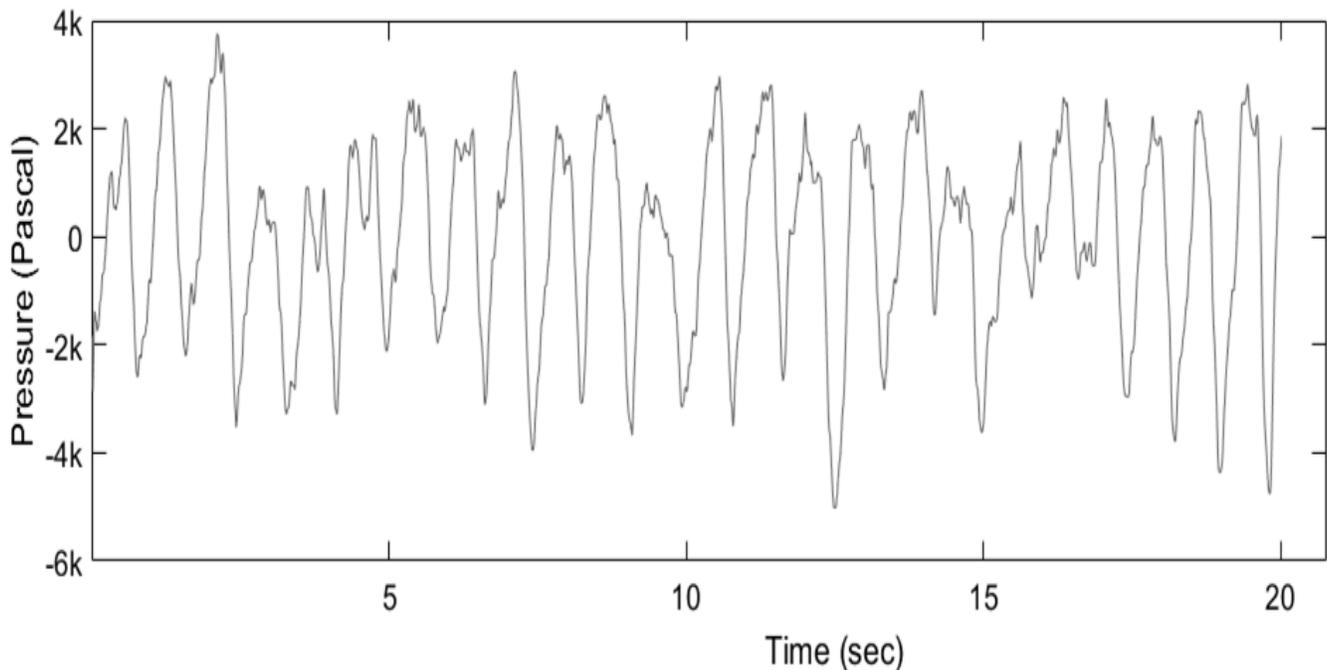


Fig. 3. Pressure signal extracted from the piezoelectric sensor voltage signal.

ACKNOWLEDGMENT

The authors wish to acknowledge the support received from Research Office at the UAE University under SURE 2016 projects grant.

REFERENCES

- [1] J. Bell, "The state of the UAE's health: 2016 | GulfNews.com," 2016. [Online]. Available: <http://gulfnews.com/gn-focus/special-reports/health/the-state-of-the-uae-s-health-2016-1.1658937>.
- [2] "The U.A.E. Healthcare Sector."
- [3] "Prospects ahead for UAE's healthcare sector – Emirates Business." [Online]. Available: <http://emirates-business.ae/prospects-ahead-for-uaes-healthcare-sector/>.
- [4] U. Anliker, J. A. Ward, P. Lukowicz, G. Tröster, F. Dolveck, M. Baer, F. Keita, E. B. Schenker, F. Catarsi, L. Coluccini, A. Belardinelli, D. Shklarski, M. Alon, E. Hirt, R. Schmid, and M. Vuskovic, "AMON: A wearable multiparameter medical monitoring and alert system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 4, pp. 415–427, 2004.
- [5] J.-C. Yu, Sung-Nien and Cheng, "A Wireless Physiological Signal Monitoring System with Integrated Bluetooth and WiFi Technologies," in *Engineering in Medicine and Biology 27th Annual Conference*, 2005, pp. 2203–2206.
- [6] Y. H. Lin, I. C. Jan, P. C. I. Ko, Y. Y. Chen, J. M. Wong, and G. J. Jan, "A wireless PDA-based physiological monitoring system for patient transport," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 4, pp. 439–447, 2004.

- [7] P. S. Pandian, K. Mohanavelu, K. P. Safeer, T. M. Kotresh, D. T. Shakunthala, P. Gopal, and V. C. Padaki, "Smart Vest: Wearable multi-parameter remote physiological monitoring system," *Med. Eng. Phys.*, vol. 30, no. 4, pp. 466–477, May 2008.
- [8] P. Bingger, J. Fiala, A. Seifert, N. Weber, A. Moser, F. Goldschmidtboeing, K. Foerster, C. Heilmann, F. Beyersdorf, P. Woias, and H. Zappe, "IMPLANTABLE MULTI SENSOR SYSTEM FOR IN VIVO MONITORING OF CARDIOVASCULAR PARAMETERS Department of Microsystems Engineering – IMTEK, Freiburg, GERMANY Department of Cardiovascular Surgery, University Hospital Freiburg, GERMANY ABSTRACT | $I R = f (ps)$," no. Fig. 1, pp. 1469–1472, 2009.
- [9] J. Im and C. Lessard, "A study for the development of a noninvasive continuous blood pressure measuring system by analyzing radial artery pulse from a wrist," ... *Med. Biol. Soc.* 1995., ..., pp. 1033–1034, 1995.
- [10] J. Kerola, V. Kontra, and R. Sepponen, "Non-invasive blood pressure data acquisition employing pulse transit time detection," *Proc. 18th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 3, pp. 1308–1309, 1996.
- [11] S. K. Kahng, "Blood-Pressure Transducer," vol. I, no. 2, pp. 54–56, 1972.
- [12] A. Giorgio, R. Diana, a. Convertino, R. Marani, and a. G. Perri, "Design of a wireless digital measurement system for the blood arterial pressure control," *2007 Asia-Pacific Microw. Conf.*, pp. 375–378, 2007.
- [13] M. C. Huang, J. C. Huang, J. C. You, and G. J. Jong, "The wireless sensor network for home-care system using XBee," *Proc. - 3rd Int. Conf. Intell. Inf. Hiding Multimed. Signal Process. IHMSP 2007.*, vol. 1, pp. 643–646,
- [14] B. A. Zneid, M. Al-zidi, and T. Al-kharazi, "Non-invasive Blood Pressure Remote Monitoring Instrument Based Microcontroller," vol. 4, no. 2, pp. 252–257, 2014.
- [15] N. Bu, N. Ueno, and O. Fukuda, "Monitoring of respiration and heartbeat during sleep using a flexible piezoelectric film sensor and empirical mode decomposition," *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 1362–1366, 2007.
- [16] D. Buxi, J. Penders, and C. van Hoof, "Early results on wrist based heart rate monitoring using mechanical transducers," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2010, pp. 4407–10, 2010.
- [17] P. Gatkine, S. Gatkine, S. Poojary, S. Chaudhary, and S. Noronha, "Development of Piezo-electric Sensor Based Non- invasive Low Cost Arterial Pulse Analyzer," *BMEiCON*, pp. 2–5, 2013.
- [18] R. Liang and Q. M. Wang, "Pulse pressure sensor based on flexible PZT thick-film composite device," *IEEE Int. Ultrason. Symp. IUS*, pp. 1559–1562, 2014.
- [19] I. Saadat, N. Al Taradeh, M. Al Ahmad, and N. Bastaki, "Non-invasive piezoelectric detection of heartbeat rate and blood pressure," *Electron. Lett.*, vol. 51, no. 6, pp. 452–454, 2015.
- [20] J. H. Park, D. G. Jang, J. W. Park, and S. K. Youm, "Wearable sensing of in-ear pressure for heart rate monitoring with a piezoelectric sensor," *Sensors (Switzerland)*, vol. 15, no. 9, pp. 23402–23417, 2015.
- [21] "Network Specifications | The XBee Alliance." [Online]. Available: <http://www.XBee.org/XBee-for-developers/network-specifications/>.
- [22] A. Pantelopoulos and N. G. Bourbakis, "A Survey onWearable Sensor-Based Systems for Health Monitoring and Prognosis," *IEEE Trans. Syst. Man, Cybern. Appl. Rev.*, vol. 40, no. 1, pp. 1–12, 2010.
- [23] S. Roundy and P. K. Wright, "A piezoelectric vibration based generator for wireless electronics," *Smart Mater. Struct.*, vol. 13, no. 5, pp. 1131–1142, 2004.
- [24] Johnson Matthey, "Datasheet Piezoceramic Masses." [Online]. Available: http://www.piezoproducts.com/fileadmin/user_upload/pdf/jm_piezoproducts_data_sheet_piezoceramic_masses_en_15_01_2015.pdf. [Accessed: 26-Sep-2016].

Performance Evaluation of a Smart Remote Patient Monitoring System based Heterogeneous WSN

Mohamed EDDABBAH, Mohamed MOUSSAOUI, Yassin LAAZIZ

Abdelmalek Essaadi University
ENSA, LABTIC laboratory
Tangier, Morocco

Abstract—This paper investigates the development of a remote patient monitoring system based on WBAN Wireless Body Sensor Network. Thus, the main purpose of such design is to interconnect heterogeneous sensor networks not equipped with the HTTP / TCP / UDP stack. A novel gateway architecture is proposed to ensure interoperability and facilitate seamless access to data from different types of body sensors that communicate via different technologies, namely, Bluetooth, IEEE802.15.4 / Zigbee and IEEE 802.15.6. Moreover, an application-layer approach for a Web Service Gateway is also developed for interaction with heterogeneous WSN. The Gateway communicates with the server via the SOAP protocol and manages the service consumption. Since the proposed platform is targeted to monitor the patient health status, a preliminary link test between the sensor and the server is unavoids in terms of quality of service. To evaluate the performances of our proposed platform, a results comparison was conducted based on different communication scenarios (3G, ADSL, and LOCAL). Finding results illustrate the (QoS) constraints, namely, Latency, Packet loss and Jitter.

Keywords—WSN; body sensor networks; remote patient monitoring; e-health; SOA

I. INTRODUCTION

To meet the growing needs of effective medical surveillance techniques, Intelligent Remote Patient Monitoring Systems (RPM) have received an important attention. Nowadays, RPMs constitute a multidisciplinary field of research such as microelectronics, telecommunications, signal processing, and computing. In the literature many RPM prototypes have been studied [1]-[3]. A Body Sensor Networks (BAN) based RPM system, several types of sensors send collected data to a server, via a communication system, and thereafter transmitted to the medical team for an eventual real-time diagnosis. However, these systems are so diverse; and integrate different communication technologies depending on the target application. In a BAN network, sensors can interact with each other and with the hub via wired or short-range wireless communication such as Bluetooth [4], IEEE802.15.4/Zigbee [5], MICS [6], ANT [7]. WBAN Networks focus on the latest advances in technology and perfectly meet the medical surveillance systems requirements. Indeed, research is conducted to identify the technical challenges in WBAN communication stack and to propose solutions [1], [8]. Recently, the IEEE802.15 workgroup has launched the IEEE802.15.6 standard dedicated exclusively to WBAN networks [9]-[11]. Remote medical surveillance

systems raise new technological challenges in terms of reliability, quality of service [12], energy consumption [13], privacy and data security [14], [15]. However a convergence between WBAN and Internet of Things (IoT) is a main key to complete these challenges and further improves the quality and efficiency of the service [16]. The biggest challenge in the design of IoT is connectivity. The creation of large-scale communicating smart object networks has become the goal of many recent and diverse research activities [4], [6]-[18]. Several systems allows direct sensors networks connection with the Internet network have been proposed [19]-[21], sensors are equipped with HTTP / TCP / UDP stack. However, connecting a simple sensor directly to Internet increases energy consumption and consequently reduces nodes life time [22], cost and complexity, especially if the sensor does not processes data. One approach is to interconnect the sensors network to Internet through a gateway that supports IP network connectivity [23], [24]. The purpose of this paper is to develop a WBAN remote patient monitoring system and interconnecting heterogeneous sensor networks not equipped with the HTTP / TCP / UDP stack, by developing a new gateway architecture to ensure interoperability and access to data from different body sensors network protocols, including: Bluetooth, IEEE 802.15.4 / Zigbee and 802.15.6, in this architecture Only the server should have a fixed IP address and just the gateway should implement TCP/IP protocols, therefore services discovery is fast, nodes management is easy, nodes coupling is low, and the design is flexible and extensible, additionally the gateway can implement more than one WBAN protocol to support heterogeneous sensors protocols. The communication between each sensor and the gateway uses a specific protocol which is the sensor protocol, however the Gateway use SOA over internet to communicate with the server. The Gateway application-layer interacts with heterogeneous WSN and avoids interferences between heterogeneous protocols. In our proposed architecture, the gateway manages the service consumption and communicates with the server via the SOAP protocol. The proposed platform targeted to monitor and process patient health status, thus platform performances evaluation is mandatory. The paper is organized as follows. Section 2 briefly describes the architecture of BAN- based RPM systems. In Section 3, we describe details on our proposed design. In Section 4, we perform several experiments to evaluate platform performances. Finally, Section 5 concludes the paper.

II. BAN- BASED RPM SYSTEMS

General architecture of a BAN-Based RPM system that corresponds to the different approaches proposed in the literature [17], [19], [25] (Fig. 1). It consists of three levels, but it may vary depending on the intended application, and may be limited to one or two levels. The first intra-BAN level consists of a BAN network that aims to centralize all the information provided by the sensors to a central coordinator, also located on the human body. An inter-BAN level provides communication between the coordinator, one or more access points located near the coordinator, or even the coordinators of other BAN networks. Finally, an extra-BAN level is responsible for routing data to a remote server via a WAN network. Data is transmitted to the medical team to obtain a real-time diagnosis or a medical database to record them, or to additional equipment for emergency alerts.

There are generally three main categories of networks in a RPM system [17]: Body Area Network (BAN), Personal Area Network (PAN), Wide Area Network (WAN). In a self-monitoring application intended to help patients to monitor their own fitness and health indicators using smart devices, smartphone or tablet; sensor nodes communicate directly with an intelligent device, forming a body area network (BAN) [26], [27]. However, in the case where the data is not processed locally but in the "cloud", the PAN network is used to connect directly, or via an access point, different BAN networks to WAN network [28]. It is also possible that the BAN coordinator sends the collected data and via WAN network to a remote server. In the BAN network, the sensors can interact with the coordinator via wired or short-range wireless (WPAN) communication such as Bluetooth, IEEE 802.15. 4 / Zigbee [24], MICS and ANT. Recently, IEEE802.15.6workgroupis established [10], which will be dedicated only to BAN applications [29]. Bluetooth and IEEE802.15.4 are two standards widely used as communication technologies between the BAN coordinator and a Network Access Point (NAP), but also as a network

gateway to cover larger areas at a lower cost. The Wi-Fi network is widely deployed as a Network Access Point which allows the WBAN network to connect to the Internet. Wide area network (WAN) systems may be cellular networks such as GSM, GPRG, 3G / 3G+ [30], LTE, Wimax, or wired communication networks (ADSL, coaxial cable and optical fiber) , or satellite networks .

III. PROPOSED REMOTE PATIENT MONITORING PLATFORM DESIGN

In this section, we propose a platform architecture for Remote Patient Monitoring systems, based on wireless sensor networks. First, we present the hardware, software platform components, and operating modes. Our RPM platform design is composed of Wireless Body Sensor Networks (WBAN) and a back-end system (Fig. 1, 2, and 3). A BAN network consists of a Gateway and a set of heterogeneous sensors. The Gateway connects the BAN to Internet; the Gateway process, storage and transfer patient data. Communication between BAN entities is an Intra-BAN communication, it is wireless and supports different communication protocols. Communication between the BAN and the server (Back-End System) is considered as an Extra-BAN communication. The extra-BAN communication can be wired or wireless via Internet using Simple Object Access Protocol (SOAP). The sensors periodically collect and send data to the gateway, which in turn transmits them to the server over Internet. Medical staff can view the patient's history after an alert or at any time. Biomedical signals can be processed locally in the BAN network or remotely in the Frond-End system. A Frond-End system supports the tracking of multiple patients, i.e. multiple BANs are served by a single Fornd-End system. The Frond-End system includes the back-end server and additional applications whose functions include the processing of received biomedical signals. The idea of outsourcing digital resources to remote servers offers very large computing and storage capabilities, unlike traditional Gateway hosting.

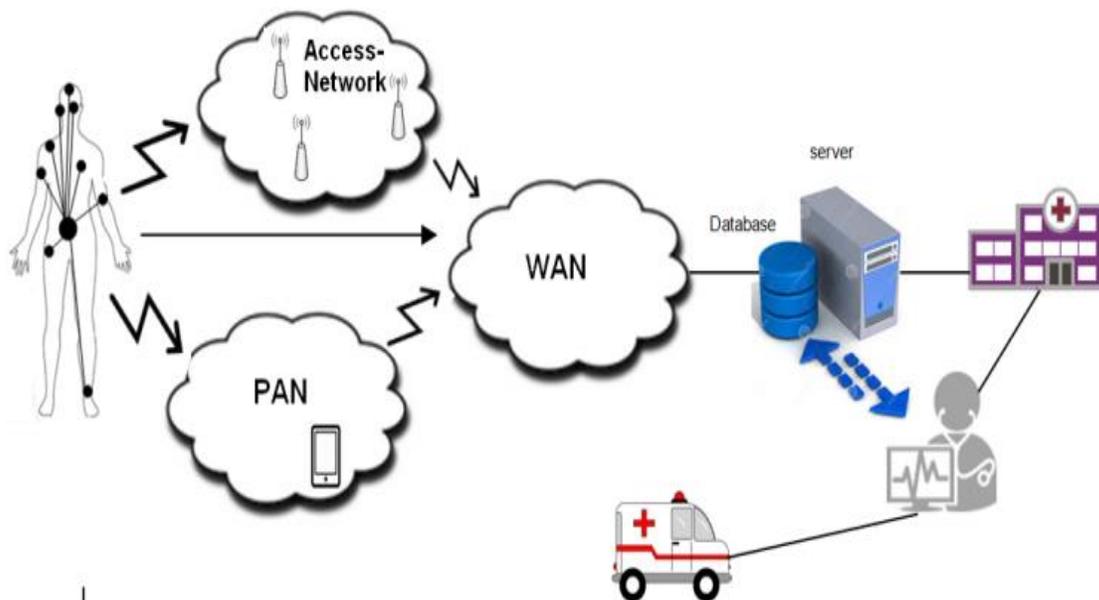


Fig. 1. General Architecture of a Body Area Network (BAN)-based RPM System.

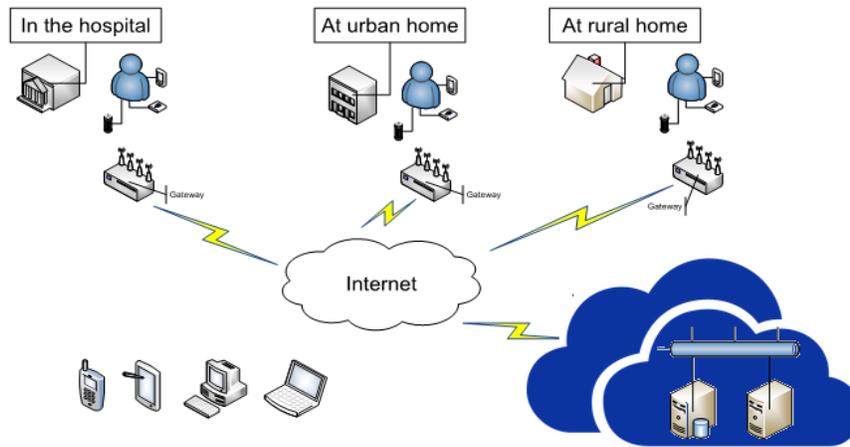


Fig. 2. Platform Architecture of ourRPM System.

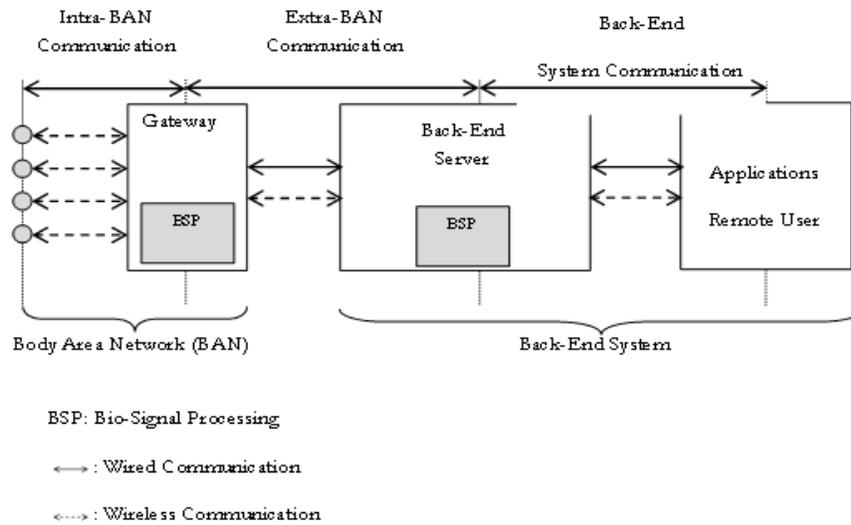


Fig. 3. General Architecture for Our Remote Patient Monitoring Platform.

A. Server

The basic services offered by the platform are previously installed on the server which has a fixed IP address, accessible from the public Internet. The Web Services Description Language (WSDL) file, provided by the server, describes the full usage of the service (methods, parameters, and return values). With the discovery service, the server detects automatically new Gateways in the network. The server database contains approved Gateways list and their attached sensors. The server receives data from the various sensors via the Gateway and stores them in a database for statistical and further examination. The server also keeps track of all transactions and medical data. It provides authentication, privacy and security. Fig. 4 shows the general protocols and services stack of the server.

B. Gateway

Gateway is developed to integrate wireless sensors deployed and web applications; it is used to connect the BAN network to Internet. The Gateway enable connectivity between

heterogeneous WSN protocols (Zigbee, Bluetooth, UWB, BAN), the gateway is completely transparent; it uses the standard communication protocols of the Web, like HTTP/HTTPS, to communicate with the Web server. XML files are used for the connection configuration, including web server address, http protocol, authentication, data priority etc.. The connection to the web server is not permanent, it is established only when the gateway wants to send data.

As shown in Fig. 5, the gateway protocols and services stack includes several PHY physical layers (Bluetooth, UWB, Bluetooth, ZigBee, WIFI ...). The application layer contains several modules that allow the gateway to manage attached nodes. The bridge layer allows the possibility to interconnect and heterogeneous sensors. The bridge layer establishes communication between two different access protocols. For example, to execute an emergency rule, a Bluetooth sensor can communicate a measurement to a ZigBee actuator. In this case every node should have several addresses like the number of supported access protocols (this information is stored in the gateway and handled by the bridge layer).

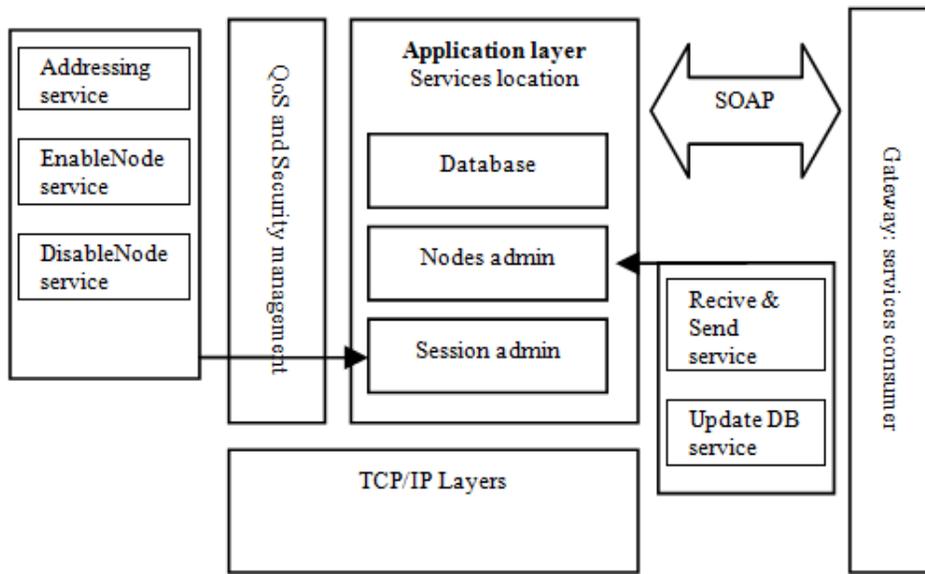


Fig. 4. Protocols Services Stack of the Server.

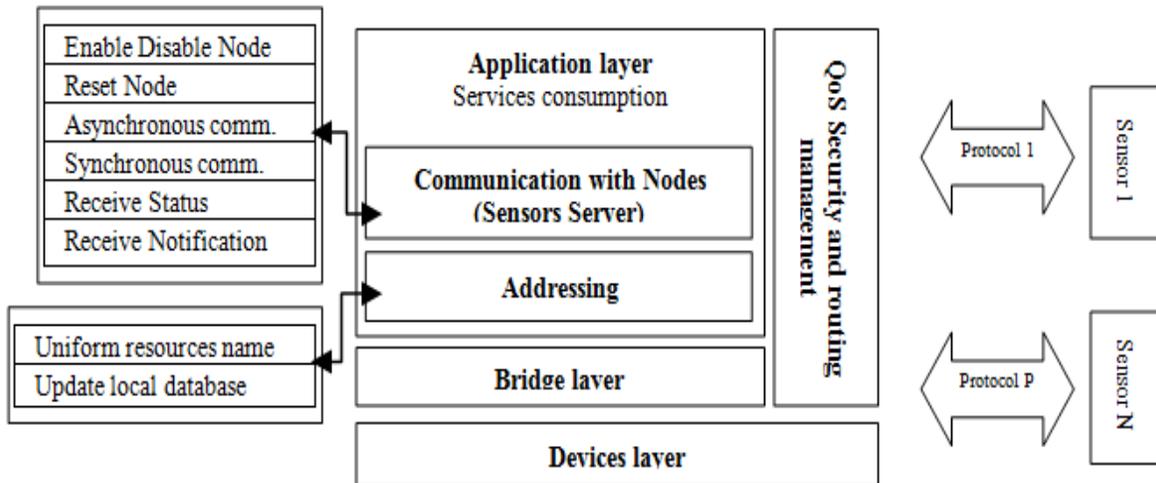


Fig. 5. Block Schematic Diagram of the Gateway.

Example of static gateway properties:

$$A \begin{pmatrix} A_{11} & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & A_{pn} \end{pmatrix}$$

A_{pn} nodes address array

P_i is an access network protocol and, $1 \leq i \leq p$

with p is the number of distinct access protocols.

N_{ij} is a P_i node, and $1 \leq j \leq n$

n is the number of nodes in the P_i network.

so A_{ij} is the N_{ij} adresse

N_{ij} is the j node in the i network

IV. PLATFORM REQUIREMENTS AND PERFORMANCES EVALUATION

The system is designed to provide continuous remote patient monitoring over heterogeneous networks. It should meet the following requirements:

- Efficient and reliable, permanent system connectivity
- Real-time Communication (RTC) services

For proof of concept, we have implemented services on remote server which communicate with the gateway via HTTP and SOAP. We have evaluated the experimental performance of the proposed platform in a real world environment under different scenarios. A typical testbed for evaluating the performance is illustrated in Fig. 6. The testbed includes three different technologies: LAN, Wifi and 3G. Here, we have used two types of test plans: request-response delay and availability test.

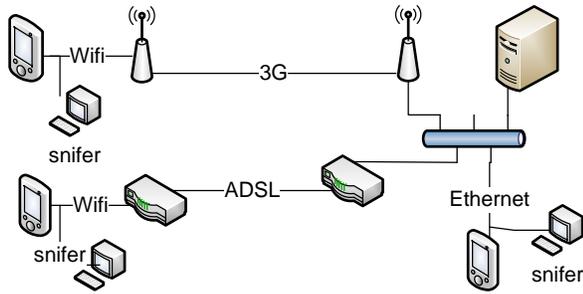


Fig. 6. Testbed.

We ran the tests with following Quality of Service (QoS) constraints: Latency, Packet loss and Jitter. We use open source packet analyzer “Wireshark” to capture packets that are being sent and received across the network.

We also ran for each scenario, a large number of request responses.

Fig. 7, 8 and 9 show the latency variation using 1000 request messages for three test scenarios: Local, ADSL and 3G.

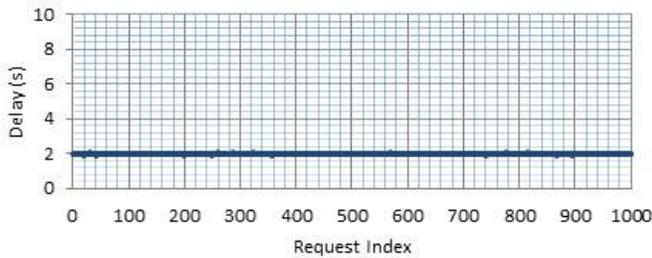


Fig. 7. Latency Variation - Local Test.

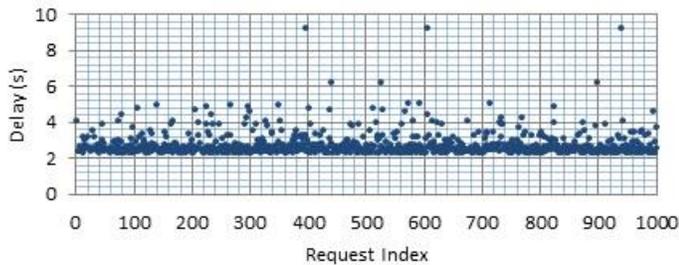


Fig. 8. Latency Variation - 3G Test.

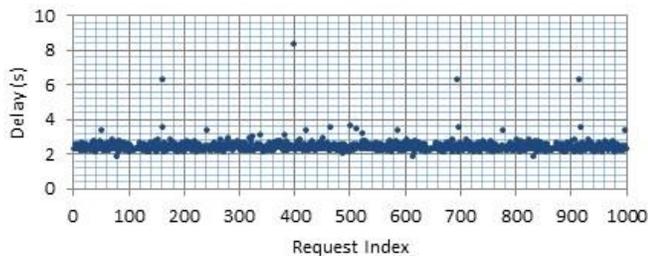


Fig. 9. Latency Variation – ADSL Test.

The test results are shown in Table I.

TABLE I. TEST RESULTS SUMMARY

	3G	ADSL	LOCAL
Number of test requests	1000	1000	1000
Successful message	864	850	977
Lost message	136	150	23
Throughput	86,4	85	97,7
Delay average (s)	2,47	2,15	2,02
Min delay (s)	2,34	1,96	1,92
Max delay (s)	10,86	8,43	2,09

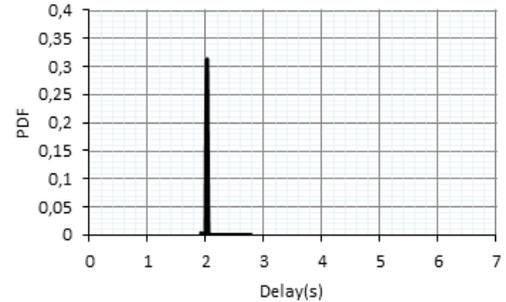


Fig. 10. Probability Distribution Function –Local Test.

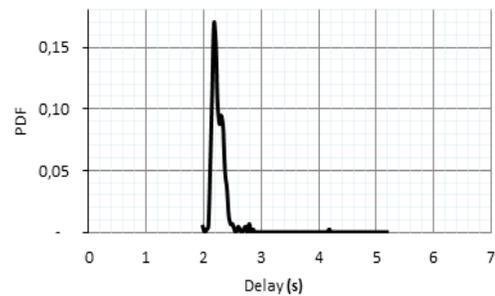


Fig. 11. Probability Distribution Function –ADSL Test.

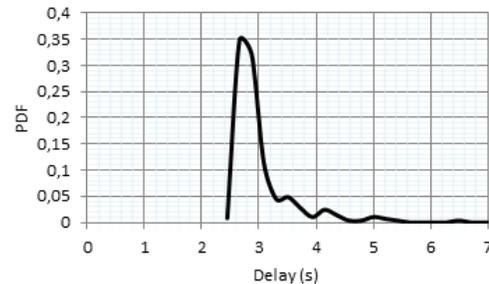


Fig. 12. Probability Distribution Function – 3G Test.

We notice that the average latency in the local test case is around 2s. This value is lower than those recorded in the two test scenarios 3G (2.47s) and ADSL (2.15s). The average latency in the 3G test is higher than those recorded in both the Local and ADSL test scenarios, this is mainly due to congestion caused by the HTTP priority configuration and network equipment. We can also observe that the packet loss rate observed during the local test equal to 2.3% is lower than those observed during 3G (13.6%) and ADSL (15%) tests. Fig. 10, 11 and 12 plot the probability distribution function of latency for the three test scenarios: Local, ADSL and 3G. As

we can see, the latency distribution takes a Gaussian form whose standard deviation depends on the jitter. High jitter means that delays are highly variable, disrupting protocols in real time. The stability of the bandwidth on the local network has meant that the jitter is significantly lower compared to the jitter resulting from the ADSL or 3G test.

V. CONCLUSION

In this paper, an application-layer approach for a WebService Based Gateway is designed. Our proposed Gateway is a communication bridge between heterogeneous sensors networks. The Gateway application layer makes sensors able to adapt and understand the various proprietary protocols. The designed Gateway uses Web standard communication protocols to communicate with the Web Server. The proposed system proves good scalability and low delay. In summary, results show that our design supports several communication modes depending on the RPM system objective (Speed, QoS, DataRate, Range).

As future steps, continued efforts are needed to complete system validation, and integration of additional sensor elements will be performed.

REFERENCES

- [1] Ullah, S., Higgins, H., Braem, B., Latre, B., Blondia, C., Moerman, I., ... & Kwak, K. S. (2012). A comprehensive survey of wireless body area networks. *Journal of medical systems*, 36(3), 1065-1094.
- [2] Alemdar, H., & Ersoy, C. (2010). Wireless sensor networks for healthcare: A survey. *Computer networks*, 54(15), 2688-2710.
- [3] Cao, H., Leung, V., Chow, C., & Chan, H. (2009). Enabling technologies for wireless body area networks: A survey and outlook. *IEEE Communications Magazine*, 47(12).
- [4] Mitra, U., Emken, B. A., Lee, S., Li, M., Rozgic, V., Thatte, G., ... & Levorato, M. (2012). KNOWME: a case study in wireless body area sensor network design. *IEEE Communications Magazine*, 50(5).
- [5] Nerino, R., Bertolo, F., Guiot, C., Bergero, D., Contin, L., & Garbin, P. (2011, May). WBSN for the Assessment of the Hippotherapy. In *Proceedings of International Conference on Body Sensor Networks*.
- [6] Yuce, M. R. (2010). Implementation of wireless body area networks for healthcare systems. *Sensors and Actuators A: Physical*, 162(1), 116-129.
- [7] Soini, M., Nummela, J., Oksa, P., Ukkonen, L., & Sydänheimo, L. (2008). Wireless body area network for hip rehabilitation system. *Ubiquitous Computing and Communication Journal*, 3(5), 42-48.
- [8] Boulis, A., Smith, D., Miniutti, D., Libman, L., & Tselishchev, Y. (2012). Challenges in body area networks for healthcare: The MAC. *IEEE Communications Magazine*, 50(5). *IEEE Communications Magazine*, vol. 50, No. 5, pp.100-106.
- [9] Kwak, K. S., Ullah, S., & Ullah, N. (2010, November). An overview of IEEE 802.15. 6 standard. In *Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, 2010 3rd International Symposium on (pp. 1-6). IEEE.
- [10] IEEE Std 802.15.6-2012 - IEEE Standard for Local and metropolitan area networks - Part 15.6: Wireless Body Area Networks
- [11] Eddabbah, M., El Ouatiqi, B., Moussaoui, M., Laaziz, Y., & Zaouiati, C. A. (2016). "Impact of BCH (51, 63, 2) code on IEEE 802.15. 6 performances". *International Journal of Computer Science and Information Security*, 14(8), 975
- [12] Chen, M. (2013, December). Mm-qos for ban: Multi-level mac-layer qos design in body area networks. In *Globecom Workshops (GC Wkshps)*, 2013 IEEE (pp. 5012-5016). IEEE.
- [13] Ababneh, N., Timmons, N., Morrison, J., & Tracey, D. (2012, March). Energy-balanced rate assignment and routing protocol for body area networks. In *Advanced Information Networking and Applications Workshops (WAINA)*, 2012 26th International Conference on (pp. 466-471). IEEE.
- [14] Hu, C., Zhang, N., Li, H., Cheng, X., & Liao, X. (2013). Body area network security: a fuzzy attribute-based signcryption scheme. *IEEE journal on selected areas in communications*, 31(9), 37-46.
- [15] Rushanan, M., Rubin, A. D., Kune, D. F., & Swanson, C. M. (2014, May). Sok: Security and privacy in implantable medical devices and body area networks. In *2014 IEEE Symposium on Security and Privacy (SP)* (pp. 524-539). IEEE.
- [16] ZAOUIAT, C. E. et LATIF, A. Internet of Things and Machine Learning Convergence: The E-healthcare Revolution. In : *Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems*. ACM, 2017. p. 62.
- [17] Custodio, V., Herrera, F. J., López, G., & Moreno, J. I. (2012). A review on architectures and communications technologies for wearable health-monitoring systems. *Sensors*, 12(10), 13907-13946.
- [18] Pantelopoulos, A., & Bourbakis, N. G. (2010). A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1), 1-12.
- [19] da Silva Campos, B., Rodrigues, J. J., Mendes, L. D., Nakamura, E. F., & Figueiredo, C. M. S. (2011, June). Design and construction of wireless sensor network gateway with IPv4/IPv6 support. In *Communications (ICC)*, 2011 IEEE International Conference on (pp. 1-5). IEEE.
- [20] Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of neuroengineering and rehabilitation*, 9(1), 21.
- [21] Colitti, W., Steenhaut, K., De Caro, N., Buta, B., & Dobrota, V. (2011, October). REST enabled wireless sensor networks for seamless integration with web applications. In *2011 Eighth IEEE International Conference on Mobile Ad-Hoc and Sensor Systems*(pp. 867-872). IEEE.
- [22] Ait Zaouiati, C. E, ENSA, I., EDDABBAH, M. M., Latif, A., & ENSA, I. (2016). "Medical Sensors Lifetime Improvement: Scheduled IEEE 802.15. 6 MAC Layer Mechanism". *International Journal of Computer Science and Information Security*, 14(8), 7
- [23] Qiu, P., Zhao, Y., Heo, U., Zhang, D., & Choi, J. (2010, June). Gateway architecture for zigbee sensor network for remote control over IP network. In *Information and Telecommunication Technologies (APSITT)*, 2010 8th Asia-Pacific Symposium on(pp. 1-4). IEEE.
- [24] Zigbee alliance (2011) Network device: gateway specification, version1.0.
- [25] Chiuchisan, I., Chiuchisan, I., & Dimian, M. (2015, October). Internet of Things for e-Health: An approach to medical applications. In *Computational Intelligence for Multimedia Understanding (IWCIM)*, 2015 International Workshop on (pp. 1-5). IEEE.
- [26] Mani, F., Makhoul, G., Oestges, C., & D'Errico, R. (2017). On the Generation of Correlated Short-and Long-Term Fading for Multiple BANs. *IEEE Antennas and Wireless Propagation Letters*, 16, 1867-1870. Barbi, Martina,
- [27] Kamran Sayrafian, and Mehdi Alasti. "Using RTS/CTS to enhance the performance of IEEE 802.15. 6 CSMA/CA." *Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016 IEEE 27th Annual International Symposium on. IEEE, 2016.
- [28] Jovanov, Emil, and Aleksandar Milenkovic. "Body area networks for ubiquitous healthcare applications: opportunities and challenges." *Journal of medical systems* 35.5 (2011): 1245-1254.
- [29] Ait Zaouiati, C. E, and Latif, A.. "Performances Comparison of IEEE 802.15. 6 and IEEE 802.15. 4". *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 8(11), 461-467 (2017)
- [30] Eddabbah, M., Moussaoui, M., and Laaziz, Y. (2014, December). A flexible 3G WebService based gateway for wireless sensor networks in support of remote patient monitoring systems. In *Proceedings of the 2014 Mediterranean Microwave Symposium (MMS2014)*, Marrakech, Morocco (pp. 12-14)

Mapping Wheat Crop Phenology and the Yield using Machine Learning (ML)

Muhammad Adnan¹

Institute of Manufacturing Information and Systems,
Department of Computer Science and Information
Engineering, National Cheng Kung University, Tainan City
701, Taiwan

Abaid-ur-Rehman², M. Ahsan Latif³, Naseer Ahmad⁴

Department of Computer Science,
University of Agriculture, Faisalabad, Pakistan

Maria Nazir⁵

Department of Computer Science,
COMSATS University Islamabad, Lahore, Pakistan

Naheed Akhter⁶

Department of Computer Science,
GC University Faisalabad, Pakistan

Abstract—Wheat has been a prime source of food for the mankind for centuries. The final wheat grain yield is the multitude of the complex interaction among the various yield attributes such as kernel per plant, Spike per plant, NSpt/s, Spike Dry Weight (SDW), etc. Different approaches have been followed to understand the non-linear relationship between the attributes and the yield to manage the crop better in the context of precision agriculture. In this study, Principle Component analysis (PCA) and Stepwise regression used to reduce the dimension of the original data to get the critical attributes under study. The reduced dataset is then modeled using the Radial Basis neural network. RBNN provides the regression value more than 0.95 which indicates the strong dependence of the yield on the critical traits.

Keywords—RBNN; PCA; stepwise regression; attributes; yield

I. INTRODUCTION

Wheat is the major agriculture crop of the Pakistan. It acts as a back bone of agriculture for the food security throughout the world. The prediction of wheat yield is too much important. The demand of the wheat has been doubled from last many decades. The demand is increasing day by day due to many factors. This may be due to many climate changes in environment. Increasing population also affect the growth and demand of the wheat. Therefore, wheat is becoming very important crop from last many years. Importance of wheat in the economy of the world is clearly reflected by its share of 15 % to the total arable land in the world for the year 2011-12 [3].

Now a days, the thing that is alarming and challenging for scientists is gap between production and the demand of wheat. Most of the people demand wheat as food .The reason is that need of wheat has been increased and it is becoming difficult to fulfill the demands. To meet such kind of challenges, it is needed to increase the total area for the agriculture land. With increasing the land area, wheat at huge quantity can be produced. There is another effort to increase the production of wheat from the present growing area.

The approach used to analyze the relation among yield and traits was machine learning (ML). Machine Learning technique has an ability to deal with high dimension problem by using less computational power. Apply machine learning in order to analyze the high number of trades to find the most relevant crop for better agriculture production. Apply machine learning algorithm for the classification of yield component in order to get high wheat yield. Principle Component Analysis (PCA) and step wise regression techniques applied on data to get the reduced dimensional data. Both techniques analyzed the data as according to its nature of effectiveness. As a result, the trimmed and the most dependent data set is achieved. New Data set collected and applied Radial Basis Neural Network (RBNN) on reduced data and got significant results. A work related to the yield measurement was conducted in which estimation of seed and grain corn yield was done on the basis of input data. ANN model with back propagation algorithm was used. The ANN model worked best with 6-4-8-1 and 6-3-9-1 structure for prediction of the yield. The result of this model is compared with multiple linear regression model. The result was approximately 95%.

The distribution of paper is following, Section 2: Related work, Section 3: Material and method, Section 4: Result and discussion, Section 5: Conclusions following with References.

II. RELATED WORKS

Adnan (2018) [1] studied the impact of water supply on wheat yield with the help of Lasso and Radial; machine learning techniques and the result of lasso Radial technique accuracy was 89% corresponding other machine learning techniques. In this study Relative water contents, waxiness, grain per spike and plant height used for experiments. Different techniques used in this study and result is clearly show that growth of wheat is highly affected by water stress. Normal-values. “Awnlength”, “pendulacnelength”, “extractionlength” and “noofdaysheading” variation is low in water stress condition as compared to Yield and TGW. Wheat yield and growth affected by water condition. In this study different techniques used for find the relationship of yield of

wheat and other variables and neural network gave the best result.

Adnan et al. (2017) [2] used the machine learning method for observation the evapotranspiration rate in Faisalabad region. In this study PCA techniques used for reduced the data set dimension because the information lost minimum with this technique. PCA gave the new variable after reduction of data set, the value of regression is 0.83426. A time series Neural Network used after getting reduced data set from PCA technique. Time series give the accurate result as compared to other Neural Network techniques. The accuracy of this model is 83%.

Awan et al., (2015) [4] described that 176 different types of genetic wheat traits were used to evaluate variety of traits practically multivariable analysis. Analysis revealed a simple correlation that indicated that there was major positive relationship of yield weight with cell membrane solidity, osmotic modification and transpiration and adverse relationship leaf area. Study also revealed significance of physical traits and their effect on grain mass. Multivariable analysis which included factor and cluster analysis showed that variable genetic pool was sufficient for breeding design. Wheat yield of each plant was strictly correlated to water substances, cell membrane solidity and leaf area. For more deep analysis, eight groups of different traits were made and study revealed that groups with smaller genetic distance were effective for breeding.

Mukhtar (2015) [5] stated that areas where major source of water is rain, has significant effect on wheat grain quality and yield because weather circumstances are randomly vary. This climate fluctuation provides chance of improving wheat grain yield production. These fluctuations and variability were studied years after year w.r.t. regions and sowing techniques and then wheat grain yield was analyzed. For this study, field tests were practiced by using three genetic traits, three different locations for the period of two years in rain source of water. Under these variations and conditions, wheat grain quality and mass resulted significant change. In region where sowing was delayed, temperature was high and water was stressed, show increased grain yield quality. However contradictory results were observed in opposing climate and water absorbing conditions. Fluctuation and variability in climate conditions had significant influence on wheat grain yield and inverse relationship was experienced among climate conditions, wheat yield and grain quality. Hence we can conclude that weather conditions, area of cultivation, sowing techniques, temperature and water can effectively alter the quality of wheat grain yield.

Emamgholizadeh et al., (2015) [6] described that in the agricultural research the most vital purpose of breeding is production of seed yield. In account to this research two techniques were used, artificial neural network and multiple regression model. Both methods were used to predict the same seed yield on the basis of premeasured features of plant like, maximum flowering days, height of the plant in centimeters, numbers of capsules of each plant and weight of seed and seed numbers. Results were tested by using both MLR and ANN techniques and it resulted that ANN was more accurate w.r.t.

root mean square error and founded coefficient. It was found also that ANN technique was better than MLR. At the end it was examined that this analysis had large and small significant effects on the same w.r.t. numbers of capsule and flower time for each plant. So in result ANN method is better for predicting seed yield than MLR and it predicts more accurately.

Khoshnevisan et al., (2014) [12] described the relation between energy consumption and crop yield in order to get sustainable agriculture they develop adaptive neuron-fuzzy system to predict wheat grain yield on the basis of energy input. The developed ANN was MLP with 11 neuron in input layer and 32 and 10 neuron in hidden layer. The result showed that ANFIS gave more accurate result than other ANN.

Paswan and Begum (2014) [7] described that how important it is for the policy maker to know about the approximate yield of crop. Computer scientists are working for making exact prediction about the yield. The crop area and crop production (maize) of Assam using ANN. They used MLP with radial bias function network which has been trained with metrological data and maize production data from various sources. The accuracy of this model was measured by using RMSE and correlation coefficient. It was observed that this model had performed better as compared to other statistical model.

Bagheri et al., (2015) [8] stated that land survey is important for crop yield prediction. Comprehensive survey may be expensive and time consuming. Since soil survey is important as it provides information for agricultural needs. Hence, there must be rapid and precise soil survey map. For this ANNs perceptron were purposed to survey map soil elements Digital Evaluation Model (DEM) features. Various multilayer ANNs were developed having input dataset and hidden element layers. This technique is implemented and tested to cumulate accuracy of interposed and inferred data. From result it was obvious that soil organization had a direct influence on accuracy of results. Errors were very small and low. Almost all techniques of ANN methods training errors were less than 11 percent. While testing and certifying, errors were 50 and 70 percent respectively. To attain superior predictions, in addition with DEM features, dataset related to lands in term of soil-farming elements must be given to ANNs perceptron as well.

Kogan et al., (2013) [9] stated that Ukraine is the biggest agricultural production country around the globe. Time management and production estimate are main elements of yield security. This study reveals wheat yield proficiency using oblast management with satellite resolution. Oblast is multinational statistics study division in European Union. Observations were made in rain fed region and average data were collected from MODIS sensing device at 250 m spatial resolution and used in a regression technique for estimating wheat yield. For reliable wheat yield projection root mean square error was acknowledged. In case of many oblasts, values which were taken in April to May using NDVI, when matched with official statistics it gave minimum root mean square error. The NDVI technique was matched with empirical model and WOFOST growth simulation applied in

CGMS, all these comparisons provided minimum RMSE. This study and comparison of wheat yield production was done totally on independent values for the period of 2010 to 2011. The most accurate forecast was predicted in 2010 via CGMS which provided root mean square error value 0.3 t ha^{-1} in June and 0.4 t ha^{-1} in April. So, it was concluded that empirical NDVI based regression was parallel to CGMS when forecasting wheat yield at oblast level.

Hung et al., (2013) [10] described that in this paper forecasting the fruit yield, multi-scale machine learning technique was used at different divisions. In this learning technique, algorithm is so flexible and usable that it can be applied at various divisions of problems. So, this approach was applied to large variety trees for fruit yield forecast. A comprehensive test was conducted on apple orchard which consisted of eight thousand images for learning. This test showed that algorithm was most fit to apple segment of various colors and sizes. Segmentation outcomes were used to count fruit and then to compare with manual counting. Squared correlation coefficient resulted from this study was $R^2=0.81$.

Alvarez (2009) [11] described that to get a model for reasonable yield prediction and grain production estimate, an analysis was conducted in Argentine grasslands in terms of soil characteristics and climate features on wheat yield. Data record collected from soil and climate analysis were implemented. Data of wheat yield production from all over the country was tested at geomorphological level. Grasslands were divided into 10 sub-regions units and from these sub-regions 10 growth seasons were recorded from 1995-2004. For data analysis, surface regression (SR) and artificial neural networks (ANNs) techniques were implemented. Yield of wheat was concentrated with water holding capacity of soil and organic carbon of soil. Climate features on yield was strong rainfall over crop potential evapotranspiration (R/CPET). Surface regression design was implemented on 64 percent of model to predict yield variance, however this design has not performed better prediction of yield. Then ANN design was tested and it gave 76 percent of yield prediction variability. So, ANN developed model was good to predict wheat yield production in Argentine grasslands.

III. MATERIAL AND METHODS

A field experiment was conducted in the University of Agriculture Faisalabad, where the growth of wheat yield was observed. That experiment was completed in two years. Where the yield was classified according to its trait values are shown in Table I. These traits were Grain Yield (GY), Kernels Per Plant (K/P), Weight/Kernel Size (KS), Number of Spikes per Plant (S/P), Number of Fertile Spikelet's per Spike (NSpt/S), Maximum Fertile Loret per Spikelet (MFFI/Spt), Spike Dry Weight (SDW), Plant Height (PH), Spike Length (SL), Awn Length (AL), Spikelets Density (SD) and Chlorophyll Contents (CC). The value of each trait was saved and was processed to find out the relation with respect to yield. In machine learning, used different approaches for classification and to find the relation of yield and variables.

TABLE I. LIST OF VARIABLES WITH ACRONYMS

GY	Grain Yield
K/P	Kernel per plant
K/S	Kernel size
NSpt / S	Number of spikes per plant
MFFI / Spt	Number of fertile spikelet per spike
SDW	Spike dry weight
PH	Plant height
SL	Spike length
AL	Awn length
SD	Spikelet's density
CC	chlorophyll contents

A. Principal Component Analysis (PCA)

PCA is a quantitatively rigorous method for achieving problem of relation. This method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. The outcome data is thus divided according to the variation in data set. The variation is done as according to the level of affecting of the data. The higher data variation is plotted first in the list of the graph. The very first plotted line represents the most variation variable. Same other plotted lines represent the gradual decreased variation of data set.

$$\text{Coeff} = \text{PCA}(X) \quad (1)$$

1) *Standardize*: We also remove unwanted data from the spread sheet data. Since PCA yields a characteristic subspace that maximizes the difference along the axes, assuming that it might have been measured on diverse scales. The conversion of the information into unit scale will be a pre requisite for those ideal executions of many machine learning algorithms.

2) *Calculate covariance*: Covariance is a measure of the degree with which components comparison is initiated for requested information move in those same heading. We find the covariance among the wheat yield and other traits. Actually we want to measure how different trait and yield depend upon each other. The formula for calculating the covariance of the variables X and Y is

$$\sum ni = 1(X - \bar{X})(Y - \bar{Y})n - 1 \quad (2)$$

With \bar{x} and \bar{y} denoting the means of X and Y, respectively. X denotes the input variable and Y denotes the output variable. Equation (2) helping in measure, how wheat yield depends upon the important variable like Kernels per Plant (K/P), Number of Spikes per plant (S/P), Number of Fertile Spikelet's per Spike (NSpt/S), Spike Dry Weight (SDW), spike length (SL), Spikelet's density (SD) that was used as input. These all above following variables are treated as Y.

3) *Selecting principal components:* That ordinary objective of a PCA is to decrease that dimensionality of the first characteristic space by projecting it onto a more abstract subspace, the place the eigenvectors will appear on those axes. However, those eigenvectors best define the directions of the new axis, since they have all the same unit length.

In this step, the PCA processes the data. The resulting value or the set of outcome which we have derived from PCA appeared in the form of eigenvector. Here, our new data after processing will be represented as eigenvector. Each principle component is a different eigenvector. The PCA has reduced the data dimension on the basis of dependency, i.e., from eleven traits into six traits. The following relation reveals the eigenvalue of an eigenvector.

$$\Sigma v = \lambda v \quad (3)$$

In equation (3):

Σ =Covariance matrix v =Eigenvector λ =Eigenvalue

To choose which eigenvector we need to drop from our lower-dimensional subspace, we must examine the relating eigenvalues of the eigenvectors. Approximately speaking, the eigenvectors for the least eigenvalues bear the slightest majority of the data over those distribution of the data and those need to be dropped. PCA provides the most significant traits by reducing the dimension of data.

4) *Transforming the samples into new subspace:* In the last step, we use-dimensional matrix W that is computed to transform our samples into the new subspace as per the following equation. The transformed new traits are used for estimating the wheat yield and equation is given below:

$$Y = W^t \times X \quad (4)$$

B. Stepwise Regression

Stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure. Stepwise regression creates a linear model and automatically adds to or trims the model. The priority in the regression model is measured according to the significant importance of the data. The data has more impact as it is added to the regression model. The data with lesser effectiveness is trimmed from the model. Only the data that is most relevant has produced targeted values whereas all the other data which is not relevant to the targeted values is discarded. This technique actually has reduced the data dimension and has given low dimensional data but highly correlated.

The stepwise model performs a multilinear regression of the response values in the n-by-1 vector y on the p predictive terms in the n-by-p matrix X. Distinct predictive terms should appear in the different columns of X with b as a p-by-1 vector of estimated coefficients for all of the terms in X. If a term is in the final model, the coefficient estimated in b for that term is a result of the final model.

C. Data Modeling

The radial basis function network is a viable alternative approach in machine learning for regression measurement in data dependency relation. A common learning algorithm for radial basis function networks is based on first choosing randomly some data points as radial basis function centers and then using singular value decomposition to solve the weights of the network. The procedure chooses radial basis function centers one by one in a rational way until an adequate network has been constructed. Here, this approach is applied on data set that was collected after obtaining the result of the PCA and the step wise regression.

In the field of mathematical modeling, a radial basis function network is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, function approximation, time series prediction, classification, and system control. Radial basis networks can require more neurons than standard feed-forward back-propagation networks, but often they can be designed in a fraction of the time it takes to train standard feed-forward networks. They work best when many training vectors are available. Here this network output layer consists of a single neuron.

IV. RESULT AND DISCUSSION

A. PCA Result

The dataset processed using PCA technique consists of 12 variables. Each variable is of different characteristics with respect to the yield production. The PCA result is shown in the Fig. 1 as each bar represents a specific principal component. The height of each bar represents the level of variation. The first component in the graph has more than 28% of the total variation in the dataset. The higher variation in a principal component reflects its significant relation with the outcome variable. The first eight principal components contribute 85% of the total variation. We can take into account this as PCA has reduced the dimension of the data by neglecting other four variables because of their least impact and variation in the graph.

Here, in Fig. 2 the graph shows PC₁ along x-axis and PC₂ along y-axis. Dependency of variables can be found out if its coefficient value is definable. From the figure, it is clear that "GY" coefficient value is higher among all other components and that is 0.46 that defines its significant role in defining the variation for the very first principal component. The trait "SDW" shows 0.41 values on the graph. Same as other variables contributing in the first component reflect their behavior from coefficient values. In component 2, which is along the y-axis the "KP" has 0.49 higher values which is

higher among all other traits. The coefficient value of “KP” in 2nd component has significant importance in 2nd principle component behavior. Similarly, it has been observed that the major contributions for PC₃ and PC₄ come from “NSpt/S” and “KS”. So, from the PCA based analysis, we concluded that the variables “GY”, “KP”, “KS”, “NSpt/S”, etc. play critical role in the final yield production.

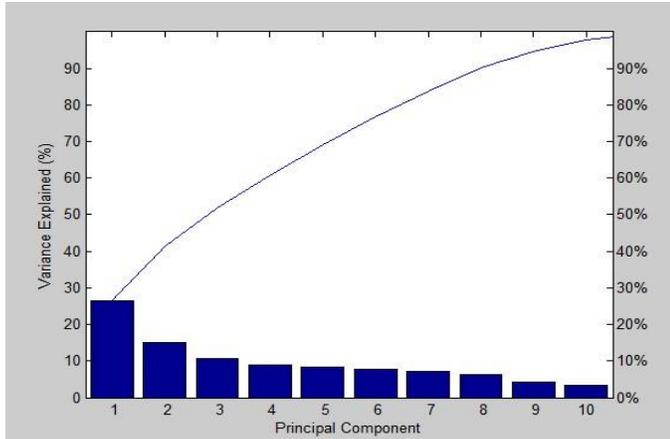


Fig. 1. PCA vs Variance.

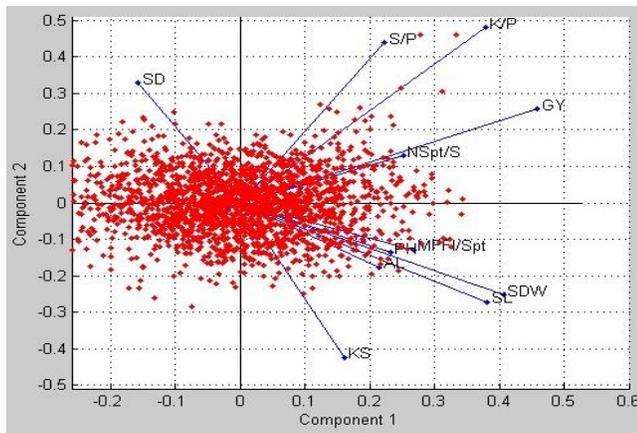


Fig. 2. PCA of the Yield and Traits.

B. Stepwise Regression Result

The stepwise regression works automatically to add or remove the predictive variables. It works best where there is large space of search. Final columns included: 1 2 3 4 6 9.

Table II explains that the model starts working when no column is added there. In the first step, it adds the kernel per plant and the value of predictive terms p is zero. In the second step, it adds the kernel size variable and the value of p =0. In the 3rd step, it adds the spike per plant variable in the stepwise model then the value of p = 2.1620e-09. In step 4, it adds the number of fertile spikelet’s per spike variable and the value of p= 0.0076. In step number six it adds the spike dry weight variable into model and the value of p= 5.2410e-06. Finally model adds Awn length, Spikelets density and Chlorophyll contents. In this process, six variables have been considered out of the total eleven variables. These six variable are “GY”, “KP”, “KS”, “SP”, “NSpt/S” and “SL”. The same variables also have been recognized by the PCA. So by the stepwise

regression we have concluded that “GY”, “KP” “KS”, “NSpt/S” show closer relation and dependency to yield in our data set. Yield production is highly dependent on these traits. This shows that stepwise also reduced the dimension of variables.

TABLE II. STEPWISE PREDICTIVE VARIABLES

'Coeff'	'Std.Err.'	'Status'	'P'
[0.0375]	[3.3429e-04]	'In'	[0]
[0.7521]	[0.0119]	'In'	[0]
[0.1126]	[0.0187]	'In'	[2.1620e-09]
[0.1044]	[0.0391]	'In'	[0.0076]
[-0.0636]	[0.1616]	'Out'	[0.6941]
[0.5198]	[0.1138]	'In'	[5.2410e-06]
[0.0113]	[0.0071]	'Out'	[0.1103]
[0.0313]	[0.0575]	'Out'	[0.5860]
[0.1517]	[0.0543]	'In'	[0.0053]
[-0.1938]	[0.2978]	'Out'	[0.5153]
[0.0042]	[0.0048]	'Out'	[0.3841]

In RBNN model trained the neural network under the data set of total eleven traits. Here, a single layered architecture is used. This model consists of 100 numbers of neurons in hidden layer and has one output layer which conventionally contained single neuron. Here radial basis function (RBF) was used as an activation function. Number of epoch in that model was 100. The obtained result from this experiment is shown in Fig. 3, the regression graph the value of regression R is 0.97695. Regression value indicate that traits and yield dependency is greater than 95%.

The model performance is best for validation value 2.6538 at Epoch number 44 which is shown below in Fig. 4. The total 44 Epoch is run by the model. The dotted line indicates the best mapping found.

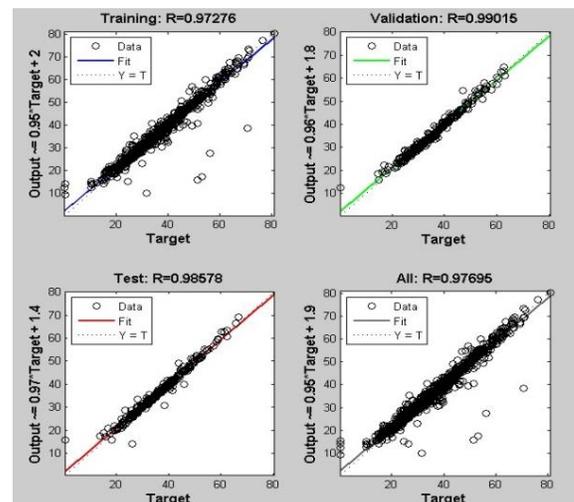


Fig. 3. Validation of Data.

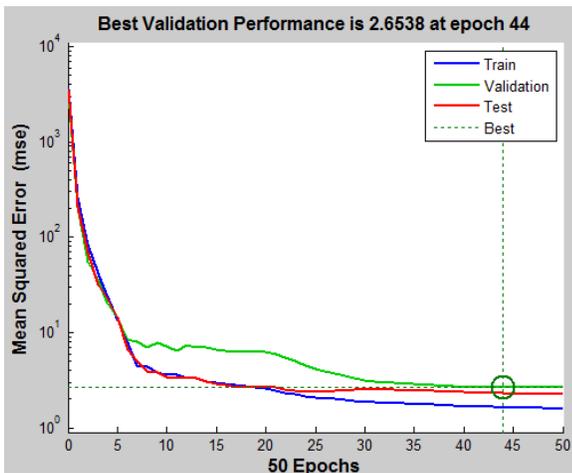


Fig. 4. Performance of Data.

V. CONCLUSION

Crop modeling is an active research area which finds its roots in the dire need to understand the mutual relationships within the crop variables. These mutual relations either linear, nonlinear or stiff in nature govern the overall crop progress and hence the yield. In this study, some of the machine learning techniques used to understand and model these relationships. The results found in this research are positive as these are highly correlated with the field results. In future work, specifically focus on the nonlinear relations which exist within these crop variables and the machine learning approaches to control that.

REFERENCES

- [1] Adnan, M., Akhter, N., Abid, M., Latif, M.A., Abaid-ur-Rehman and Kashif, M., 2018. Studying the Impact of Water Supply on Wheat Yield by using Principle Lasso Radial Machine Learning Model. *International journal of advanced computer science and applications*, 9(2): 229-235
- [2] Adnan, M., Latif, M.A. and Nazir, M., 2017. Estimating Evapotranspiration using Machine Learning Techniques. *International journal of advanced computer science and applications*, 8(9): 108-113.
- [3] Khan, M.U., Malik, R.N. and Muhammad, S., 2013. Human health risk from heavy metal via food crops consumption with wastewater irrigation practices in Pakistan. *Chemosphere*, 93(10):1-8.
- [4] Awan, S.I., Ahmad, S.D., Ali, M.A., Ahmed, M.S. and Rao, A., 2015. Use of multivariate analysis in determining characteristics for grain yield selection in wheat. *Sarhad J. of Agric.*, 31: 139-150.
- [5] Ahmed, M., 2015. Response of spring wheat (*Triticum aestivum* L.) quality traits and yield to sowing date. *PLoS one* 10(4):40-56.
- [6] Emamgholizadeh, S., Parsaeian, M. and Baradaran, M., 2015. Seed yield prediction of sesame using artificial neural network. *European Journal of Agronomy*, 68: 89-96.
- [7] Paswan, R.P. and Begum, S.A., 2014, February. ANN for prediction of Area and Production of Maize crop for Upper Brahmaputra Valley Zone of Assam. *In Advance Computing Conference (IACC), 2014 IEEE International* : 1286-1295.
- [8] Bagheri Bodaghabadi, M., Martínez Casasnovas, J.A., Salehi, M.H., Mohammadi, J., Esfandiarpour Borujeni, I., Toomanian, N. and Gandomkar, A., 2015. Digital soil mapping using artificial neural networks and terrain-related attributes. *Pedosphere* 25 (4): 580-591.
- [9] Kogan, F., Kussul, N., Adamenko, T., Skakun, S., Kravchenko, O., Kryvobok, O., Shelestov, A., Kolotii, A., Kussul, O. and Lavrenyuk, A., 2013. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models. *International Journal of Applied Earth Observation and Geoinformation*, 23: 192-203.
- [10] Hung, C., Underwood, J., Nieto, J. and Sukkarieh, S., 2015. A feature learning based approach for automated fruit yield estimation. *In Field and Service Robotics, Springer* : 485-498.
- [11] Alvarez, R., 2009. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. *European Journal of Agronomy* 30(2): 70-77.
- [12] Khoshnevisan, B., Rafiee, S., Omid, M. and Mousazadeh, H., 2014. Development of an intelligent system based on ANFIS for predicting wheat grain yield on the basis of energy inputs. *Information processing in agriculture* 1(1): 14-22.

It has been observed that by applying PCA, we have got the results with greater accuracy. In this way, we have reduced the computational time and power by using reduced and new variables provided by the PCA. The reduced variables provide almost same results as we have got from all the available variables to map the yield.

We also have applied some different techniques and methods in this neural network model to have different regression values as shown in Table III that help to find out the best relation among the traits and yield.

TABLE III. DATA MODELING ANALYSIS

Stepwise Regression	Sr#	Input			Output
		No. of Neuron	Activation Function	Training Function	
	1	30	Logsig	trainlm	0.967
2	40	Tansig	trainlm	0.949	
PCA	3	30	Logsig	trainlm	0.955
	4	40	Tansig	trainlm	0.943

Soft Error Tolerance in Memory Applications

Muhammad Sheikh Sadi, Md. Shamimur Rahman, Shaheena Sultana, Golam Mezbah Uddin, Kazi Md. Bodrul Kabir
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh

Abstract—This paper proposes a new method to detect and correct multi bit errors in memory applications using a combination of a clustering approach, Bit-Per-Byte error detection technique, and Majority Logic Decodable (MLD) codes. The likelihood of soft errors accelerates with system complexity, reduction in operational voltages, exponential growth in transistor per chip, increases in clock frequencies, breakdown of memory reliability and device shrinking. Memories are the sensitive part of a computer system. Soft errors in memories may cause an instruction to malfunction. Several techniques are already in practice to mitigate the soft errors. Majority logic decodable codes are proved as effective for memory applications because of their ability to correct a massive number of errors. Since memories are used to hold large number of bits that's the restraint of Majority logic decodable codes method, so we emphasize on the size of data word in this method. The proposed method aims to detect and correct up to seven bit errors with lesser computational time. It works in an efficient manner in case of adjacent errors which is not possible in Majority logic decodable codes (MLD). It is delineated by Experimental reviews that the proposed approach outperforms existing dominant approach with respect to number of erroneous bit detection and correction, and computational time overhead.

Keywords—Soft error tolerance; bit-per-byte; majority logic decodable codes; clustering; adjacent errors

I. INTRODUCTION

The unusual condition of multifaceted nature, and the way that the software and hardware are so unpredictably connected, denotes that the system might be extremely delicate to soft errors. In particular, soft errors are a matter of great concern when planning high accessibility systems or systems utilized as a part of electronic-antagonistic situations [1]-[4]. In memory applications, soft error can change an instruction or any data value [3]-[5]. Almost all system chips have embedded memories like ROM, DRAM, SRAM, flash memory etc. But soft errors in such memory applications are increasing alarmingly as technology these days is focusing on smaller dimension of devices which leads to the integration of circuits [6]. Integrated circuits are prone to particle strike or radiation which can cause the memory cell to change its state and obtain a different value than what was desired. Small size of transistors, capacitors and low operating voltages are also the reasons for soft error in memories. So, fault tolerant technique in memory architecture is fundamental issue to ensure its reliability to the users. A small flaw or glitch in a memory cell can change an instruction or can cause a whole program to work incorrectly leading to inappropriate information or loss of valuable data.

There are some existing dominant approaches to provide fault tolerance in memory applications. For example, for satellite applications, hamming code and parity codes are used to secure memory devices. There are some other methods for error detection and correction such as Error Correction Code (ECC) [7]-[9], Euclidean geometry low-density parity check (EG-LDPC) codes [10], [11], etc. However, almost all of these methods are facing area, and time overhead, and significant power consumption penalty. Also these methods have low error detection and correction rate and exhibits lower performance while working with large data word. To overcome these barriers, we came up with a fault tolerant technique which can work with larger data word and consume lesser processing time.

In this paper, an error detection and correction technique is proposed to protect the memory applications. This method combines the salient features of clustering approach [12], Bit-Per-Byte error detection technique, and Majority Logic Decodable (MLD) codes [13]-[16]. Majority Logic Decodable codes are used because of their ability to detect multiple bit upsets; Bit-per-byte technique minimizes the required time to detect the error; and the clustering approach works in a very efficient manner in case of adjacent errors. The proposed method provides high efficiency for error detection and correction and can correct up to 7-bit upsets in a 49-bits' data block.

The rest of this paper is presented as follows. Section 2 provides several related work in this area of research. The proposed methodology and associated examples are discussed in Section 3. Experimental analysis is shown in Section 4. Section 5 concludes the paper.

II. RELATED WORK

First Several techniques are already in practice to provide error detection and correction. Some of them are discussed below.

Naeimi et al. [8] proposed a fault-tolerant memory architecture which can tolerate faults both in the storage unit and in the encoder or decoder. A fast and compact error correcting technique is proposed in that paper which is known as one step majority logic correction. One step majority logic correction works in a way that it corrects every erroneous bit at each step and will output the correct code word after full processing. This method requires the same number of cycles as the number of bits for both detection and correction which is a major degrade in performance in terms of access time in memory.

Shih-Fu et al. [7] presented an error detection method for different set cyclic codes using majority logic decoding scheme. Majority logic decodable codes are most appropriate for memory applications because they deal with large number of errors but it may lower the memory performance with excessive decoding time. MLD was first introduced for Reed-Muller codes. They described a plain majority logic decoder (MLD) whose circuit arrangement includes four components: i) a cyclic shift register; ii) an XOR matrix; iii) a majority gate; and iv) an XOR for correcting the code word bit under decoding. It can correct multiple bit-flips depending on the number of parity check equations [6]. They proposed a modified version of MLD which is known as Majority Logic Detector/Decoder. The MLDD technique needs 15 cycles to correct an error. However, it can detect and correct only two bit errors from a 15-bit data word and the time requirement of this method is high enough to degrade its performance in terms of access time in memory.

Jayalakshmi et al. [5] came out with a modified representation of MLDD. It overcomes the existing techniques by detecting errors in lesser cycles. They used additional logic which results in an area overhead. Another limitation is that this method needs additional three cycles to correct any error.

III. PROPOSED METHODOLOGY TO DETECT AND CORRECT ERRORS

In this chapter, the proposed method will be discussed and explained elaborately. The chapter will take you step by step through our method to have a better understanding about the method. Some examples along with pictorial representation will be provided with the method explanation.

A. Memory with MLDD

The existing MLDD [5] is modified to improve its performance. Euclidean Geometry Low Density Parity Check Codes (EG-LDPC) [6] works behind the existing MLDD. The following Fig. 1 shows how the MLDD modification proposed by us will be used in a memory system.

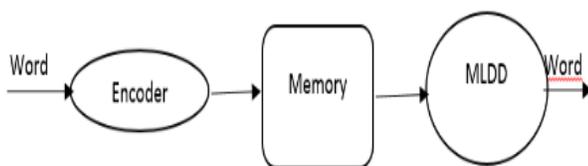


Fig. 1. Proposed Structure of a Memory System with MLDD.

B. Encoder Architecture

The design of encoder is generated from the EG-LDPC codes. The following parameter are in the function of EG-LDPC for any integer $t \geq 2$, where t is the number of errors that the code can correct.

- Information bits, $k = 22t - 3t$
- Code word Length, $n = 22t - 1$
- Minimum distance, $d_{min} = 2t + 1$

Let's consider $t=2$ and if the other parameters are determined accordingly then we would have a (15, 7, 5) EG-LDPC code which will have a generator matrix like Fig. 2 and if Fig. 3 the architecture of an encoder circuit [7] for (15, 7, 5) EG-LDPC code is shown. The information bits are indicated from $i_0 \dots i_6$. The check bits are calculated using linear sum (XOR) operation of the information bits. The information bits are copied to the encoded vector from $c_0 \dots c_6$ and the check bits are copied from $c_7 \dots c_{14}$. Thus the encoded matrix is generated.

	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}
i_0	1	0	0	0	0	0	0	1	0	0	1	1	1	0	1
i_1	0	1	0	0	0	0	0	1	1	0	0	1	1	1	0
i_2	0	0	1	0	0	0	0	0	1	1	1	0	0	0	1
i_3	0	0	0	1	0	0	0	1	0	1	1	1	0	0	0
i_4	0	0	0	0	1	0	0	0	1	0	1	1	1	1	0
i_5	0	0	0	0	0	1	0	0	0	1	0	1	1	1	0
i_6	0	0	0	0	0	0	1	0	0	0	1	0	1	1	1

Fig. 2. Generator Matrix of (15, 7, 5) EG-LDPC code [8].

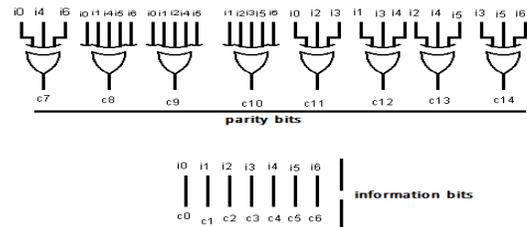


Fig. 3. Architecture of an Encoder Circuit for the (15, 7, 5) EG-LDPC code.

C. Design Structure of Corrector

One-step majority-logic is a fast and efficient error-handling technique [10]. There is a class of ECCs that are one-step-majority correctable. Type-I two-dimensional EG-LDPC is one of the example of one-step-majority correctable codes. In this section, the one-step majority-logic corrector for EG-LDPC codes is shown.

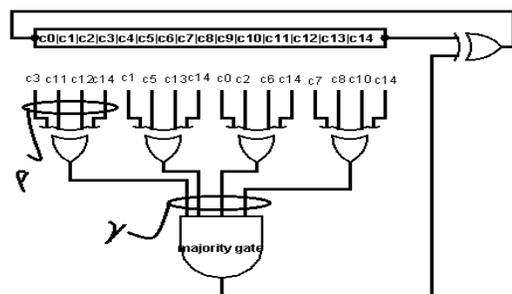


Fig. 4. Serial One-Step Majority Logic Structure to Correct Last Bit (Bit 14th) of 15-bit (15, 7, 5) EG-LDPC code [8].

A linear sum named Parity-Checksum can be formed by computing the internal product of the received vector and a row of a parity-check matrix. The principle of the one-step majority-logic corrector is generating parity-check sums from the defined rows of the parity-check matrix. These steps correct a potential error in one bit e.g., c_{n-1} .

1) Generate parity-check sums by calculating the inner product of the received vector and the defined rows of parity-check matrix.

2) The check sums are fed into a majority gate. If the output of majority gate is “1”, then the bit c_{n-1} is corrected by inverting the value of c_{n-1} .

The architecture of a serial one-step majority logic corrector for (15, 7, 5) EG-LDPC code is shown in Fig. 4.

D. Fundamental Concepts of Proposed Methodology

The proposed methodology uses the MLDD [5] technique described above as a part of correction method. Our proposed method is tested for a 49-bit data block and it can correct up to 7 bit errors. We proposed a clustering idea to divide consecutive seven bit placed in different cluster. That’s why this proposed method can be applied where there is need to detect and correct adjacent multiple cell upset (MCU). Because adjacent bits are in different cluster and change in adjacent bits can detect easily and correct. The method is discussed below:

1) At first the data word which has the size of 49 bit, is clustered into 7 clusters keeping distance 7 between the data bits or information bits. We will keep 7 bits in each cluster. So this will result in $49/7=7$ clusters. Now each cluster will have the information as shown in Fig. 5. The 49 data bits are represented as $a_1, a_2, a_3, \dots, a_{49}$. Then form 7 different clusters such as $a_1, a_8, a_{15}, a_{22}, a_{29}, a_{36}, a_{43}$ and adjacent bits like a_1, a_2, a_3 are placed in different clusters.

2) Each cluster has 7 information bits. Now we calculate even parity for each cluster. It is quite similar to the idea of bit-per-byte technique. If we consider each cluster as a byte (although each cluster here has 7 bits and a byte is formed of 8 bits) then we can apply the bit per byte technique on the clusters like a bit-per-cluster. We have used even parity technique here to assign parity to the clusters. Even parity means the number of 1’s must be even. If number of 1’s is even then parity is 0, otherwise parity is 1 to make number of 1’s is even. So after this step, each cluster has it corresponding parity which will be sent with the information bits. We can visualize it as shown in Fig. 6.

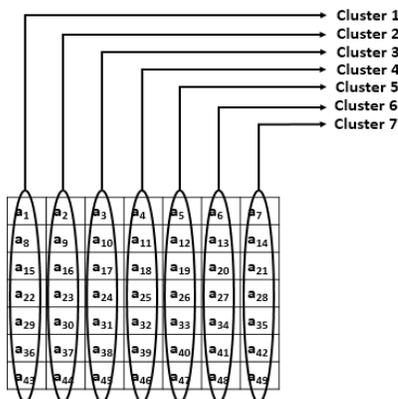


Fig. 5. Architecture of Seven Clusters with 49 Information Bits.

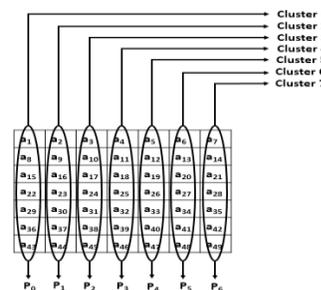


Fig. 6. Calculated Parity Bits for Each Cluster.

3) Now we are going to apply Majority Logic Detector (MLDD) scheme for each cluster. Let’s consider each cluster has information bits denoted as i_0, \dots, i_6 . Then according to the MLD [7] we have generated the check bits from the information bits which are the checksums (XOR) of information bits. The check bits are generated as shown in Fig. 7.

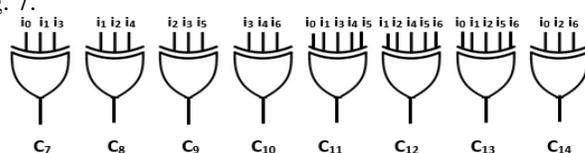


Fig. 7. The Architecture to Generate Check Bits.

Now the clusters have 7 information bits and 8 check bits which is 15 bits.

4) In this step, the information bits will be sent to the receiving side in the form which was seen at the first step like $a_1, a_2, a_3, \dots, a_{49}$. With the information bits, parity bits of each cluster will also be sent which was calculated using odd parity. Along with these, the check bits for each cluster are also sent to the receiving end. So, the following information are sent from the sending end.

- Information bits ($a_1, a_2, a_3, \dots, a_{49}$)
- Parity bits for each cluster ($p_0, p_1, p_3, \dots, p_6$)
- Check bits generated for each cluster ($C_7, C_8, C_9, \dots, C_{14}$)

5) This information is sent to the receiving side. While transmitting the above information, any bit may flip and change the state from 0 to 1 or 1 to 0 resulting in misleading information. At the receiving end the information bits will be received but they may not be error free. Let the received information bits are $a_1, a_2, a_3, \dots, a_{49}$

6) At the receiving end, we will form clusters like we did in step 1. So we will have 7 clusters keeping distance as 7 among the information bits of each cluster. Finally, the generated clusters are- Cluster1, Cluster2, Cluster3, ..., Cluster 7.

7) After forming the clusters, we will calculate the parity bits for each cluster using odd parity. So the parity of each cluster at the receiving end may look like- parity (Cluster1), parity (Cluster2), parity (Cluster3) ... parity (Cluster7).

8) In this step parity of each cluster of sending end will be compared with the parity of receiving end's cluster. If a mismatch is found at any cluster, then that cluster will be taken under consideration and that cluster is assumed to have error in its bits. Now let's assume Cluster (i) have a mismatch and it has errors. Now check bits will be generated for that cluster using the technique as described in step 3. So after generating the check bits ($C_7, C_8, C_9, \dots, C_{14}$) we will have total 15 bits to apply the majority logic decoding. The information bits are copied to C_0, C_1, \dots, C_6 . So the code word will be like: $C_0, C_1, C_3, \dots, C_{14}$.

9) The process of majority logic decoding is outlined shortly as follows:

Step 1: Initialize counter variable to 0.

Step 2: Calculate majority values B_j as follows:

$$B_1 = C_3 \oplus C_{11} \oplus C_{12} \oplus C_{14} \quad \text{Eq. (1)}$$

$$B_2 = C_1 \oplus C_5 \oplus C_{13} \oplus C_{14} \quad \text{Eq. (2)}$$

$$B_3 = C_0 \oplus C_2 \oplus C_6 \oplus C_{14} \quad \text{Eq. (3)}$$

$$B_4 = C_7 \oplus C_8 \oplus C_{10} \oplus C_{14} \quad \text{Eq. (4)}$$

Step 3: If majority value is greater than 2 then go to step 4, else go to step 5.

Step 4: Inverse the 14th bit. Store the counter which is the erroneous bit position. Go to step 5

Step 5: Perform one-bit cyclic left shift.

Step 6: Increment the counter

Step 7: If counter variable equals to 8 then go to step 8 else go to step 2

Step 8: End

10) Now we have the positions where bit flip in a cluster has occurred during transmission and those erroneous bits are corrected. We store those positions in a cluster to determine the actual positions in the data word. Next we examine other clusters to find errors (if any) and find their positions in the corresponding cluster and thereby correct them. If we follow this method, then we would be able to detect and correct adjacent bit upsets which is a common issue in memory applications. Let's walk through an example to describe our method with sending end code word of Fig. 8 and receiving end code word of Fig. 9. Sending code word is the original data with parity bits and receiving code word is the erroneous collection of original code word.

For the above example, total seven clusters can be formed with the above forty-nine data bits. Now, the parity bits of receiving clusters are compared with those of the sending clusters.

0	1	1	1	0	0	0
1	1	0	0	1	0	0
1	1	1	0	1	1	1
0	0	0	0	1	0	0
0	1	0	0	1	0	0
0	1	1	0	1	0	1
0	1	1	1	0	0	0

Fig. 8. Sending Code Word.

0	0	1	1	0	0	0
1	1	0	0	1	0	0
1	1	1	0	1	1	1
0	0	0	0	1	0	0
0	1	1	0	1	0	1
0	1	1	0	1	0	1
0	1	1	1	0	0	0

Fig. 9. Receiving Code Word.

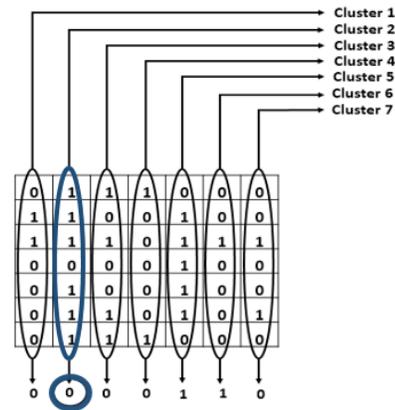


Fig. 10. Parity Bits of Sending Part.

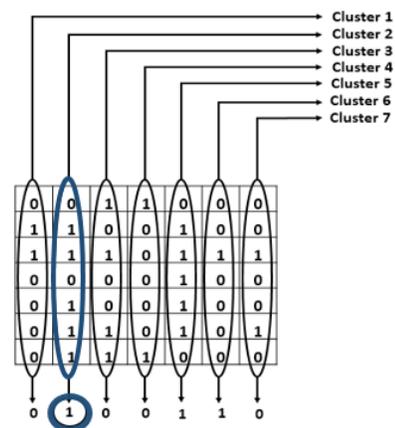


Fig. 11. Parity Bits of Receiving Par.

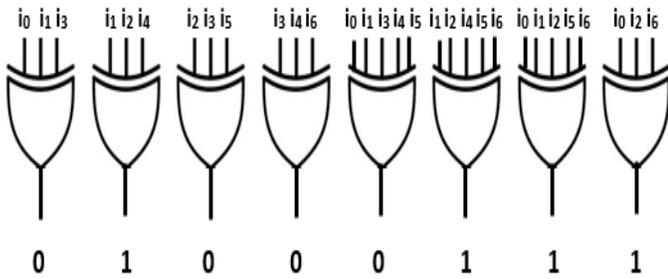


Fig. 12. Calculate Check Bits when Mismatching in Sending and Receiving Parity Bits.

If there is any mismatch, then only for this cluster we will generate 8-bit parity using Majority Logic Detector Decoder (MLDD) scheme.

As shown in Fig. 10 and 11, we can observe that in second cluster there is a mismatch and for this cluster we will generate 8-bit parity using the following architectures shown in Fig. 12.

Then for the erroneous cluster, the size of the code word will be 15-bit. i.e. $C_0, C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}, C_{12}, C_{13}$ and C_{14} . In this case, it will be 011011101000111.

Using majority decoding circuit, we will perform eight left cyclic shift. At each step of shift operation, the majority values $B_1, B_2, B_3,$ and B_4 will be calculated. If the majority values are 1 then it is confirmed that the current bit under decoding is erroneous. Then an inverter is added to the 14th bit position in the register. The whole procedure of eight cycles is shown in Fig. 13.

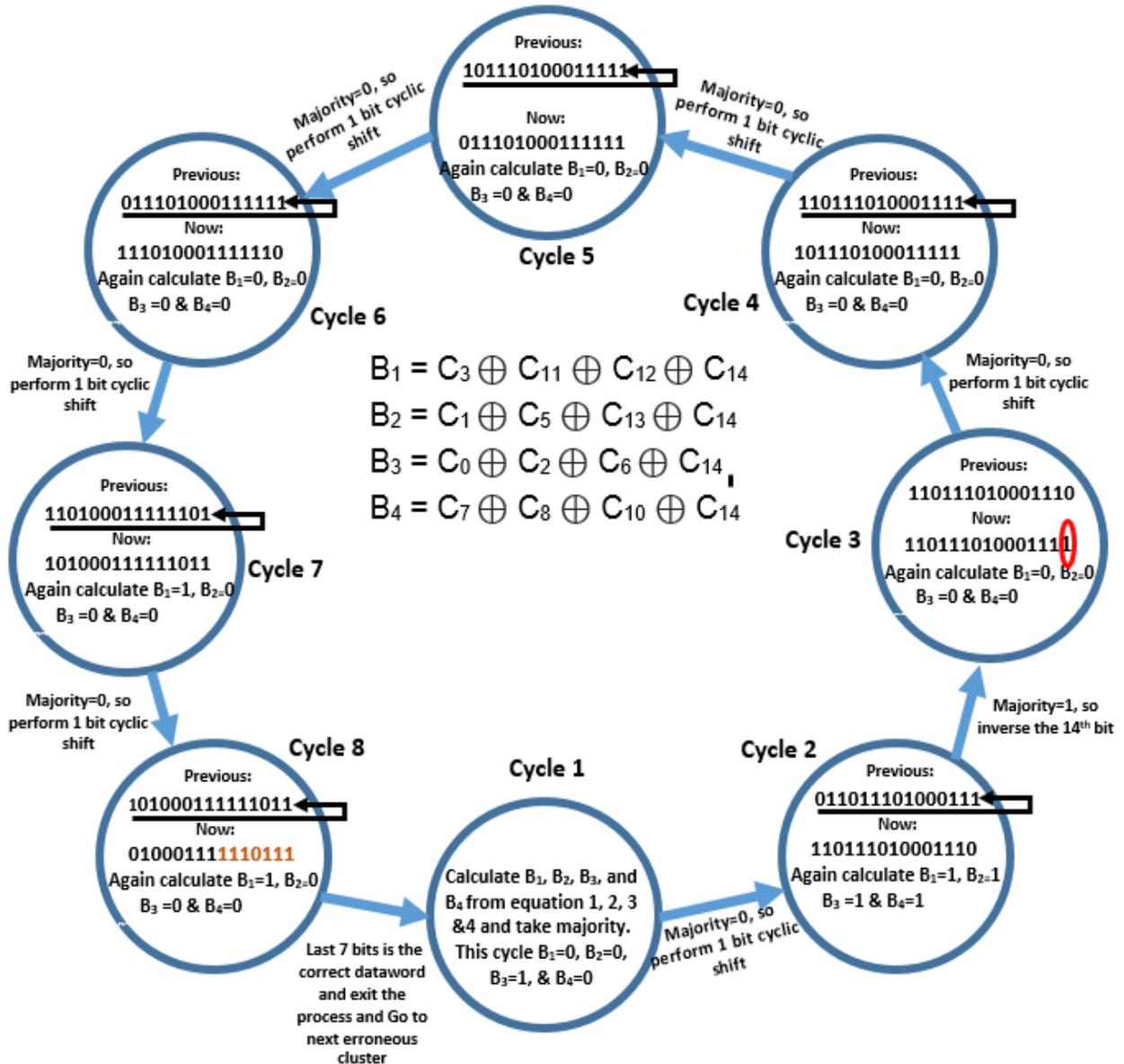


Fig. 13. Performing Eight Left Cyclic Shift for Acquiring the Error Free Code Word.

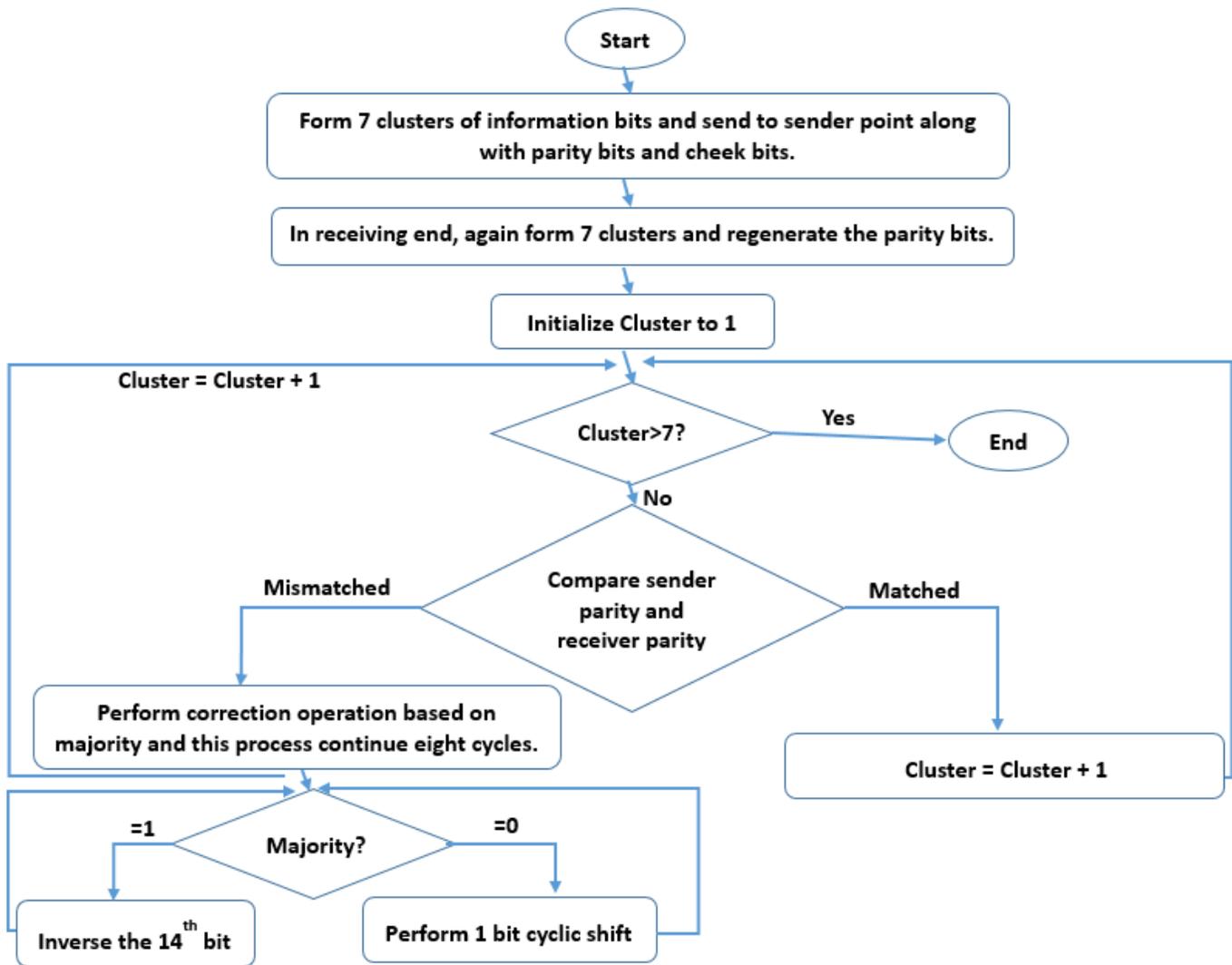


Fig. 14. Flow Chart of the Proposed Method.

In cycle 1, calculate B_1, B_2, B_3 and B_4 using the above (1), (2), (3) and (4). Then check the majority and this cycle we get $B_1=0, B_2=0, B_3=1$ and $B_4=0$. So majority is 0 and performs one bit cyclic shift and goes to cycle 2. The values of B_1, B_2, B_3 and B_4 are again calculated and this time majority is 1. So according to the proposed algorithm, the 14th bit is inverted and goes to cycle 3. This procedure is repeated till cycle 8 with the two possibilities, one is majority 0 then perform one bit cyclic shift and another is majority 1 then inverse the 14th bit.

After the 8th cycle we can see the original 7 information bits are in last 7 position. Hence, if we do seven right shift then we will get the corrected code word

The corrected code word is: 1 1 1 0 1 1 1 0 1 0 0 0 1 1 1.

After going through the whole process, we will get original information bits as expected to be received. Then from the clusters we obtain the information bits of the form $a_1, a_2, a_3, \dots, a_{49}$. Now the overall workflow of the proposed method is

shown in Fig. 14 as a flow chart which provides a better overview of the method.

IV. EXPERIMENTAL ANALYSIS

This proposed methodology is experimented through a simulation procedure. The simulation process includes 'error-detection' phase and 'error-correction' phase. It identifies the soft error through the detection phase and appropriately recovers it so that the original stored data is retrieved. In this section, the experimental results of proposed method and other existing methods are represented and discussed. The effectiveness of the proposed method is evaluated in this section.

A. Experimental Tools

The following tools are used for the evaluation process of the proposed method.

- Intel(R) Core i5-2430M CPU @ 2.40 GHz
- CPU RAM 6GB

- Language: Python 3.4
- IDE PyCharm 5.0.1

B. Experimental Result

The outcomes of the experiments are shown in this section along with some comparisons with the existing methods. The results ultimately indicate how the proposed method performs better in terms of the amount of cyclic shift needed. Also it shows that the proposed method performs better to deal with common mode errors or adjacent bit errors while the existing methods are not suitable for this purpose. Fig. 15 shows the comparison of cycle needed for error detection by the plain MLD [8] and existing MLDD [5], and the proposed method.

In all cases MLD [8] occupy 15 cycles to detect errors. In case of MLDD [5], if there is no error then it takes only three

cycles to confirm that one. But if there is error, then it takes larger cycles. However, the proposed method requires fewer cycles than MLD [8] and MLDD [5] to detect any error for 14-bitcode word using bit per byte and clustering approach.

Fig. 16 shows the comparison of cycle needed for error correction by the plain MLD [8], existing MLDD [5] and the proposed method.

If an error is detected, MLD takes 15 cycles need to run the entire decoding process. The existing MLDD needs 18 cycles. The existing MLDD has same procedure. However, rather than 15 cycles, three additional cycles are required. The proposed method needs $(15+3)/2$ cycles that means 9 cycles.

If two-bits error are detected, MLD [8] needs 15 cycle for correction. MLDD [5] needs $(15+3)$ cycles that means 18 cycles but the proposed clustering method it needs 16 cycles.

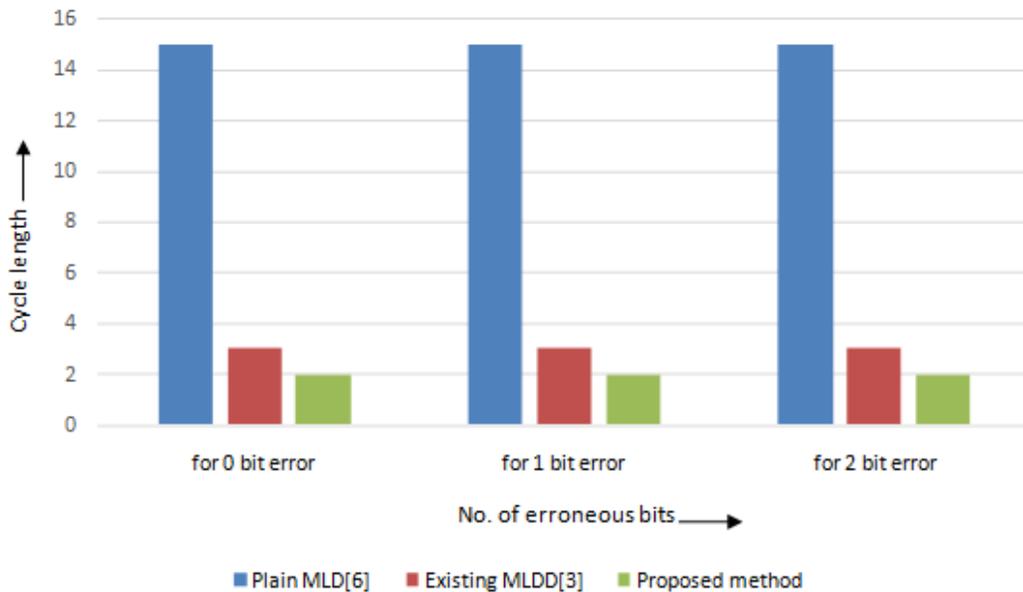


Fig. 15. The Comparison among Plain MLD [6], the Method Proposed by Jayarani et al. [3], and the Proposed Method for Error Detection.

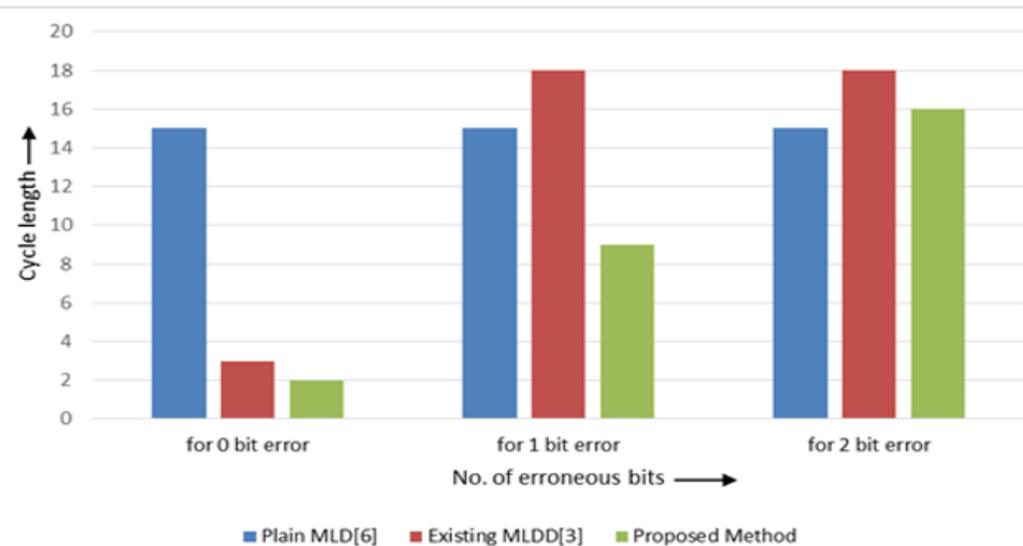


Fig. 16. The Comparison among Plain MLD [8], the Method Proposed by Jayarani et al. [5], and the Proposed Method for Error Correction.

V. CONCLUSIONS

The proposed methodology focuses on the architecture of a Majority Logic Decoder/Detector (MLDD) with the utilization of bit-per-byte and clustering approaches for fault detection and correction, with decreased cycles. Along with this, the proposed method is very much useful when there are errors in adjacent bits because each adjacent bit is formed in different cluster. So that errors can be easily detected. So, those systems where much possibility to occur adjacent bit error then this proposed method perform better than any other MLDD system with minimum cycle. The proposed method is designed in a way so that it could deal with larger data block. Experiments are performed for large data word to prove its efficiency. To show better performance with larger data block our clustering based approach may consume more time than other methods which are good for smaller data word. The proposed method can detect and correct multiple adjacent cell upsets whereas, the existing cannot perform that. The main limitation is that when multiple errors occur in same cluster then the proposed method can't detect these faulty bits. This proposed method is only focused to detect adjacent error and minimum cycle than the existing. In the later work, we try to detect and correct errors in same cluster and work with large data block quite faster than this proposed method.

REFERENCES

- [1] Shanshan Liu , Jiaqiang Li, Pedro Reviriego , Marco Ottavi, and Liyi Xiao "A Double Error Correction Code for 32-Bit Data Words With Efficient Decoding" in IEEE TRANSACTIONS ON DEVICE AND MATERIALS RELIABILITY, VOL. 18, NO. 1, MARCH 2018.
- [2] Jiaqiang Li, et al, "Efficient Implementations of 4-Bit Burst Error Correction for Memories" IEEE Transactions on Circuits and Systems II: Express Briefs.
- [3] J. Yang et al., "Radiation-Induced Soft Error Analysis of STT-MRAM: A Device to Circuit Approach," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 35, no. 3, pp. 380-393, March 2016.
- [4] Jing Guo; Liyi Xiao; Tianqi Wang; Shanshan Liu; Xu Wang; Zhigang Mao, "Soft Error Hardened Memory Design for Nanoscale Complementary Metal Oxide Semiconductor Technology," IEEE Transactions on Reliability, vol.64, no.2, pp.596,602, June 2015.
- [5] K.Jayalakshmi, B.Sivasankari, "Reduction of Decoding Time in Majority Logic Decoder for Memory Applications", "International Journal of Innovative Research in Computer and Communication Engineering", Vol.2, Special Issue 1, March 2014.
- [6] R.Meenaakshi Sundhari, C.Sundarrasu, M.Karthikkumar, "An Efficient Majority Logic Fault Detection to reduce the Accessing time for Memory Applications", "International Journal of Scientific and Research Publications", Volume 3, Issue 3, March 2013.
- [7] Shih-Fu Liu, Pedro Revingo, and Juan Antonio Maestro "Efficient majority fault detection with difference set codes for memory applications", IEEE Trans. Very Large Scale Integration. (VLSI) Syst., vol. 20, no. 1, pp. 148–156, Jan. 2012.
- [8] H. Naeimi and A. DeHon, "Fault secure encoder and decoder for Nano Memory applications," IEEE Trans.Very Large Scale Integration. (VLSI) Syst., vol. 17, no. 4, pp.473–486, Apr. 2009.
- [9] R.C.Baumann,"Radiation-induced soft errors in advanced semiconductor technologies," IEEE Trans. Device Mater.Rel. , vol. 5, no.3, pp. 301–316, Sep. 2005.
- [10] C. W. Slayman, "Cache and memory error detection, correction, and reduction techniques for terrestrial servers and workstations,"IEEE Trans. Device Mater. Reliabil., vol. 5, no. 3, pp. 397–404, Sep. 2005.
- [11] Y Kato and T. Morita, "Error correction circuit using difference-set cyclic code," Proceedings of the ASP-DAC Asia and South Pacific Design Automation Conference, 2003.
- [12] Costas A. Argyrides, Pedro Reviriego, Dhiraj K. Pradhan and Juan Antonio Maestro "Matrix-Based Codes for Adjacent Error Correction" IEEE Transaction on Nuclear Science (Vol. 57 No.4), August 2010.
- [13] Pedro Reviriego, Juan A. Maestro, and Mark F. Flanagan, "Error Detection in Majority Logic Decoding of Euclidean Geometry Low Density Parity Check (EG-LDPC) Codes,"IEEE Transactions On Very Large Scale Integration (Vlsi) Systems 1.
- [14] R. J. McEliece, The Theory of Information and Coding. Cambridge,U.K.: Cambridge University Press, 2002.
- [15] R. Lucas, M. P. C. Fossorier, Yu Kou, Shu Lin, "Iterative decoding of one-step majority logic deductible codes based on belief propagation", IEEE Transactions on Communications (Volume:4 Issue: 6), June 2000.
- [16] Y. Kou, S. Lin, M. P. C. Fossorier, "Low-density parity-check codes based on finite geometries: a rediscovery and new results", IEEE Transactions on Information Theory (Volume:47 , Issue: 7), Nov 2011.

Safety and Performance Evaluation Method for Wearable Artificial Kidney Systems

YeJi Ho¹, SangHoon Park², KyungMin Jo³,
Barum Choi⁴, SangEun Park⁵

Asan Institute for Life Sciences, Asan Medical Center,
Seoul, Korea

Jaesoon Choi*⁶

College of Medicine, University of Ulsan, Ulsan, Korea,
Asan Institute for Life Sciences, Asan Medical Center,
Seoul, Korea

Abstract—This paper focuses on international standards and guidelines related to evaluating the safety and performance of wearable dialysis systems and devices. The applicable standard and evaluation indices for safety and performance are determined, and the relevant international standards and guidelines are provided in a table. In addition, example experiments using a triaxial accelerometer and robot arm are presented for testing the endurance and safety of wearable artificial kidneys. The findings in this paper can be used to suggest new guidelines for the mechanical safety and performance evaluation of wearable artificial kidney systems.

Keywords—Wearable artificial kidney; safety; hemodialysis; peritoneal dialysis; accelerometer

I. INTRODUCTION

As lifestyle-related metabolic disorders such as diabetes and hypertension continue to proliferate with westernized diets, the number of patients with end-stage kidney failure has drastically increased. When end-stage kidney failure is caused by diabetes, which is most frequently the case, the 5-year survival rate is only 50%–60%. This is much lower than the rates for stomach cancer and colorectal cancer. In the USA, approximately 45,000 people died of end-stage kidney failure in 2006; it is ranked as the ninth-leading cause of death among Americans.

According to the organ transplant status provided by the Korea Organ Network for Organ Sharing (KONOS) and US Department of Health & Human Service, the number of patients waiting for kidney transplants soared between 2000 and 2015, but actual cases of kidney transplants increased only slightly. Because the supply is far less than the demand for kidney transplants, the majority of patients are treated with dialysis therapy. In the medical device industry, the hemolysis-related market reached 75 billion USD in 2011 and is continuing to grow each year.

Until now, the hemolysis-related market has been focused on supplying equipment and consumables. However, total renal care service—integrating the complete process of hemolysis including management, service, and remote therapy—is expected to dominate the future market. A representative example of this trend is the self-dialysis unit for home use.

Around the world, about 11% of patients treated with peritoneal dialysis and 0.6% of those treated with hemolysis have been reported to stay at home for therapy. This indicates the rapid growth and increasing use of home-use self-dialysis units [1], [2]. The entire market related to artificial kidney and hemolysis is expected to rapidly grow.

New types of wearable artificial kidneys for patients with end-stage kidney failure are being actively developed [3]–[9], however there are no criteria available for evaluating the safety and performance of wearable artificial kidneys. In this study, existing local and overseas standards and guidelines were analyzed, and evaluation indices were derived. A safety and performance evaluation method is proposed for new types of wearable artificial kidneys.

II. STANDARDS FOR SAFETY AND PERFORMANCE EVALUATION

A. New Types of Wearable Artificial Kidneys

In most advanced countries, portable artificial kidney devices are classified as wearable or home-use types according to their purpose and place of usage. Home-use artificial kidney devices share common characteristics, but small artificial kidney devices are portable. Accordingly, the latter group was set as a single category, and the following three types were considered in this study: wearable, house-use and portable.

Artificial kidneys can be classified as house-use or portable types and have been already developed for self-dialysis or peritoneal dialysis at home. In addition, safety and performance evaluation systems for these artificial kidneys have been established. Thus, this study considered wearable artificial kidneys that can be used outside the hospital without the help of medical experts. With regard to common matters such as physiological functions, performance, and major safety considerations, the corresponding provisions in existing standards for other types of artificial kidneys (e.g., IEC, ISO, KS) were applied.

B. International Standards and Guidelines Applicable to Wearable Artificial Kidneys

Because there are no standard and guidelines for wearable artificial kidneys, international standards and guidelines for hemodialysis/peritoneal dialysis system and portable medical devices were evaluated if they seemed applicable to wearable artificial kidneys. Table I presents the international standards and guidelines.

This research was supported by grants (No. 15172MFDS434 and No. 17172MFDS340-4) from Ministry of Food and Drug Safety in 2015 and 2017.

TABLE I. APPLICABLE STANDARDS AND GUIDELINES FOR WEARABLE ARTIFICIAL KIDNEY SYSTEMS

	Standard No.	Item
International Standards	IEC 60601-2-16	Hemodialysis
	IEC 60601-2-39	Peritoneal dialysis
	IEC 60601-1-6	Usability
	IEC 60601-1-11	Home healthcare environment
	IEC 62366-1	Usability engineering
	IEC 60529	Degrees of protection provided by enclosure (IP code)
	IEC 60721-4-7	Portable and nonstationary medical devices
	ANSI/AAMI RD5	Hemodialysis systems
	Guidelines and National Standards	State (Regulation authority)
South Korea (MFDS)		29 artificial kidney system
USA (FDA)		Guidelines for industry and FDA staff: hemodialysis blood tubing sets
		Guidelines for hemodialysis delivery systems
Europe (EU)		EN 60601-2-16, EM 60601-2-39
Japan (PMDA)		JIS T 0601-2-39, JIS T 3250
China (CFDA)		YY 0053-2008, YY 0054-2010

TABLE II. EVALUATION INDICES FOR SAFETY AND PERFORMANCE

Evaluation indices		Contents	Related standards
Safety	Transit: operable and portable	Vibration	Broadband vibration test
		Battery	Check the backup power and indication of state
		Push	Enclosures of ME equipment shall have sufficient rigidity to protect against unacceptable risk
		Molding stress relief	Enclosures of molded or formed thermoplastic materials shall be constructed so that any shrinkage or distortion of the material is due to the release of internal stresses
		Shock	Shock test
	Alarm	Alarm condition and overlap mode	Normal operation under each condition
	IP code	First characteristic numeral (hazardous parts)	Protection against access to hazardous parts
		First characteristic numeral (solid foreign objects)	Protection against solid foreign objects
		Second characteristic numeral (water)	Test for second characteristic numeral 2 with the drip box
Performance	Dialysis fluid temperature	Measurement of the dialysis fluid temperature	

The international standards for hemodialysis and peritoneal dialysis systems are IEC 60601-2-16 and IEC 60601-2-39, respectively. In this study, new types of wearable artificial kidneys were defined as medical devices used by patients, not medical experts. Accordingly, IEC 60601-1-11 should be applied for the home healthcare environment, including

medical devices for home use and point of care. IEC 60601-1-6 and IEC 62366-1 should be applied for usability.

Apart from international standards, relevant guidelines and national standards were also surveyed. Both the Food and Drug Administration (FDA) and Ministry of Food and Drug Safety (MFDS) provide guidelines for hemodialysis based on IEC

60601-2-16. Especially in the case of medical devices for the high-risk group, the innovation pathway of FDA allows some documents required for license to be submitted after the devices come to the market. This allows for a fast examination to issue a license. No wearable artificial kidneys have yet been licensed. In 2012, Gura et al.'s [3] wearable artificial kidney was the first to be permitted for clinical application. In Europe, Council Directives 93/42/EEC and 90/385/EEC for medical devices established the EN-60601-x standard based on the IEC 60601 series and enforce its observation.

In Japan (PMDA) and China (CFDA), the Japanese Industrial Standards and YY (medicine and medical device standard) of China Compulsory Certification are based on IEC 60601-2-16 and IEC 60601-2-39 and include additional requirements for matters such as special test methods.

III. EVALUATION INDICES

Common applicable indices for safety and performance evaluation were derived from the surveyed national standards and guidelines that seemed to be relevant to wearable artificial kidney. These are presented in Table II.

The safety evaluation indices included characteristics of transit-operable and portable devices, which are specified by IEC 60601-1-11 (home healthcare environment), alarm systems (IEC 60601-1-8), and waterproof/dustproof-related safety. A common performance evaluation index is the dialysis fluid temperature, but it appeared to differ by country and standard, so it was excluded from the evaluation indices in this study.

IV. EXAMPLE EVALUATION METHODS

The above safety and performance evaluation indices are based on existing standards for dialysis systems. Environmental tests such as vibration and shock tests for transit-operable and portable devices, which are specified in IEC 60601-1-11, do not seem to sufficiently consider the intrinsic features of wearable artificial kidneys. Accordingly, example methods for evaluating the durability and safety that consider the device features are presented below. A wearable artificial kidney being developed by a research team at Seoul National University was subjected to the evaluation methods. Because the artificial kidney is not completely developed, a dummy without internal dialysis circuits and sensors was used. A triaxial accelerometer experiment and trajectory experiment using a robot arm were applied as the example methods for evaluating the durability and safety. The triaxial accelerometer experiment was performed to identify the acceleration range acting on a device during activities of daily living such as walking, running, and climbing up and down stairs. The trajectory experiment utilized a robot arm for various trajectory movements such as vertically reciprocal, circular, and falling motions.

A. Triaxial Accelerometer Experiment

1) *Purpose:* As shown in Fig. 1, a subject wore the wearable artificial kidney around his or her waist. A triaxial accelerometer was used to detect the range of acceleration that could affect the wearable artificial kidney system during

movements of daily living (e.g., walking, running, and climbing up and down stairs).

2) *Method:* Nine adult subjects wore a wearable artificial kidney and performed easy activities of daily living (walking, running, and climbing up and down stairs. As shown in Fig. 2, a triaxial accelerometer (MMA7260Q, Freescale Inc., Austin TX) was used to measure the acceleration acting on the dummy according to the movement of the hip joint. A short-distance wireless transmitting and receiving module (Bluetooth 2.0, OpenbrainTech Inc., South Korea) was attached to transmit acceleration signals to the PC. Each subject wearing the artificial kidney dummy with inserted sensors conducted 10 rounds of 25 m reciprocal walking and running movements. In addition, the subjects also climbed up and down 13 stairs 10 times (Fig. 3). The acceleration was measured for the movements of the wearable artificial kidney dummy in the directions of the mediolateral, anteroposterior, and superoinferior axes (Fig. 4). The measured accelerations were saved with LabVIEW2013 (NI Inc., Austin, TX). The measurements were subjected to a fast Fourier transform (FFT) to identify vibrations of the dummy during each movement in the frequency domain.

3) *Results:* When the subjects wearing the artificial kidney dummy walked and climbed up and down stairs, the frequency distribution was concentrated between 0.9 and 2.1 Hz on average. When the subjects wearing the artificial kidney dummy ran, the frequency distribution was usually concentrated at 3 Hz and above (Table II, Fig. 5).



Fig. 1. Wearable Artificial Kidney System Dummy.

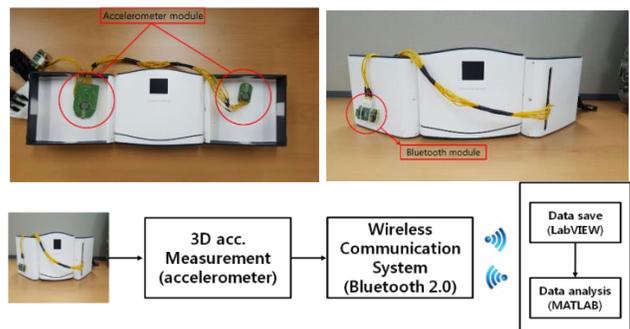


Fig. 2. Measurement System for the Triaxial Acceleration.

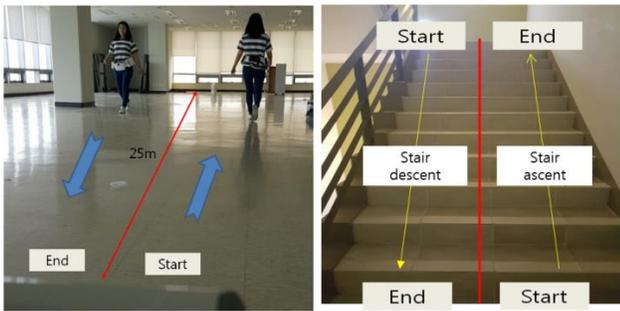
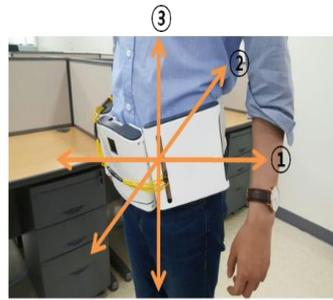


Fig. 3. Experimental Setup for Activities of Daily Living.



- ①: Medialateral axis
- ②: Anteroposterior axis
- ③: Superoinferior axis

Fig. 4. Measurement Directions for the Triaxial Acceleration.

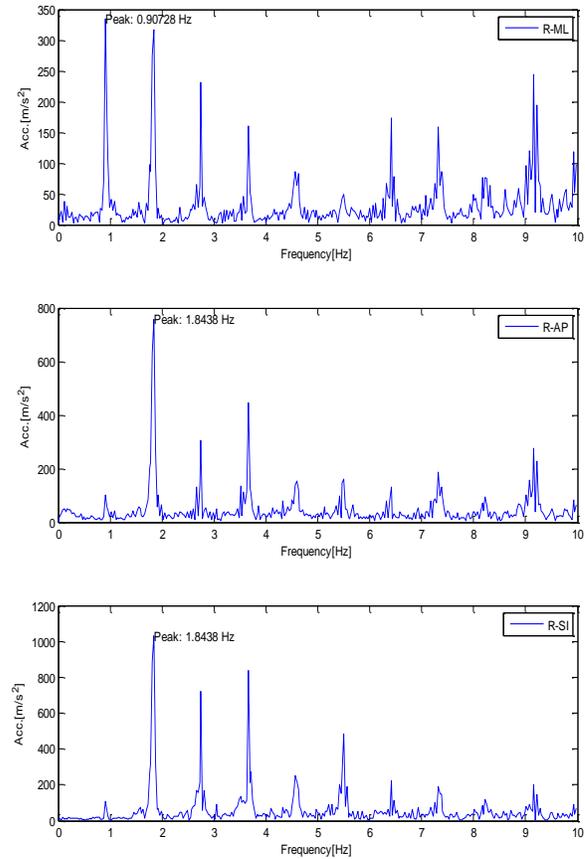
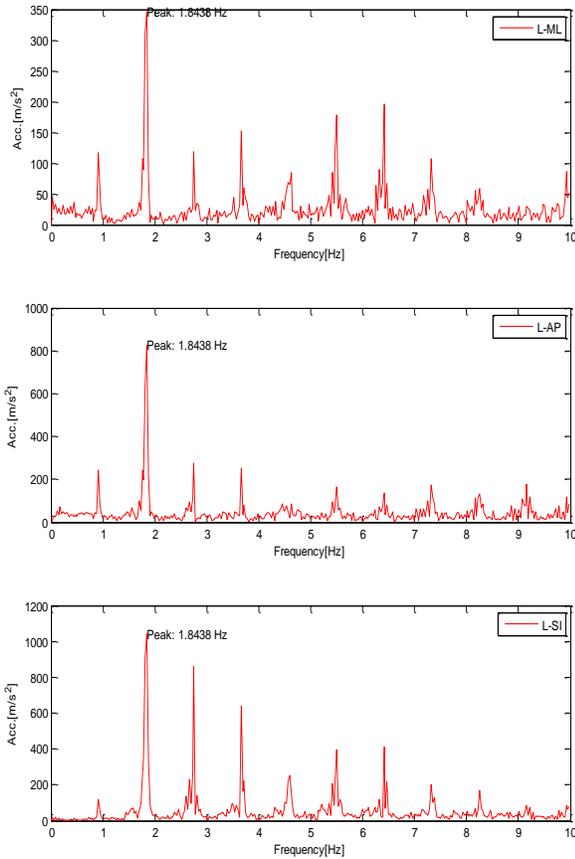


Fig. 5. Frequency Domain During Activities of Daily Living: Walking.

B. Robot Arm Experiment

1) *Purpose:* An experiment using a robot arm was performed to judge the normal operation of the wearable artificial kidney system during and after various trajectory movements like reciprocating vertical/circular/falling motions are repeated.

2) *Method:* As shown in Fig. 6, the wearable artificial kidney dummy was fixed to the robot arm (Hyundai Inc., Korea). The experimental conditions were set for vertical stroke, circular, and falling motions.

V. CONCLUSION

In this study, existing local and international standards and guidelines for dialysis systems and artificial kidneys were surveyed to propose a safety evaluation method for wearable artificial kidneys. The surveyed standards and guidelines were used to derive safety and performance evaluation indices that seem applicable to wearable artificial kidneys. Along with the evaluation methods specified by existing standards, example methods for evaluating the durability and safety are presented. A triaxial accelerometer or robot arm was applied to evaluate an artificial kidney dummy being developed by a Korean research team.

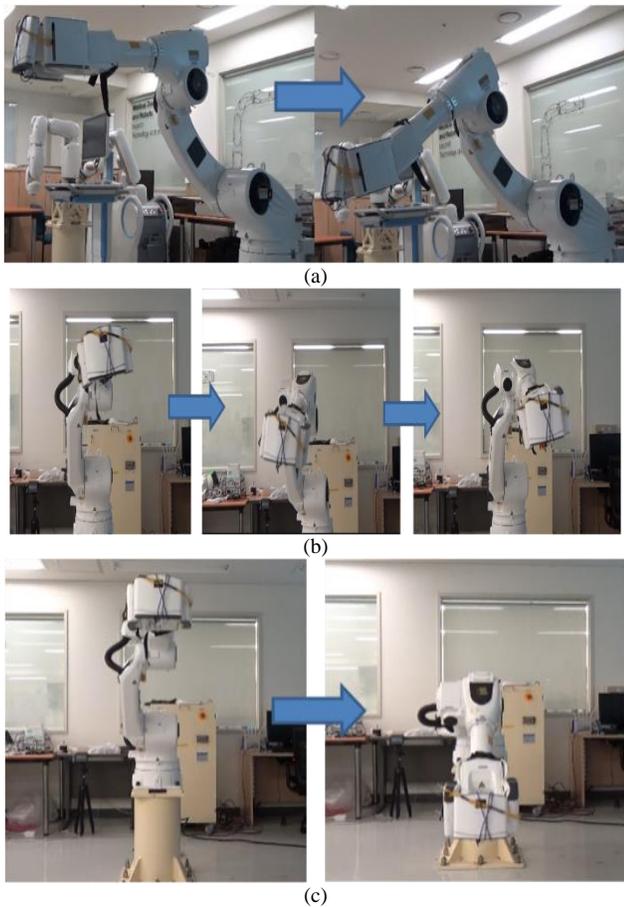


Fig. 6. Experimental setup for the robot arm trajectory: (a) stroke vertical motion (range: 400 mm, speed: 270 mm/s), (b) circular motion (range: 400 mm (CCW), speed: 270 mm/s), and (c) falling motion (range: 1 m, speed: 1.5 m/s).

Because the artificial kidney dummy does not include any internal dialysis circuits it cannot be considered a real artificial kidney system. Accordingly, even if the dummy was evaluated with the presented methods, its normal operation and safety could not be identified. Thus, the evaluation methods for durability and safety need to be applied to a prototype wearable

artificial kidney that includes dialysis circuits and sensors. Future studies also need to check the normal operation and safety of a real wearable artificial kidney after evaluation. In addition, it should be performed additional daily activities such as sitting, lying down and falling [10]-[12], which were not performed in this study. If a motion capture system is used to measure the motion of the hip joint during each movement and the measurements are reflected in the trajectory experiment with a robot arm, the range of normal operation will be defined more effectively.

REFERENCES

- [1] S. J. Shin, "Current status of bioartificial kidney," *J. Biomed. Sci. Eng.*, vol. 7, pp. 108–111, 2004.
- [2] J. C. Kim and C. Ronco, "Current technological approaches for a wearable artificial kidney," *Contrib. Nephrol.*, vol. 171, pp. 231–236, 2011.
- [3] V. Gura, A. S. Macy, M. Beizai, C. Ezon, and T. A. Golper, "Technical breakthroughs in the wearable artificial kidney (WAK)," *Clin. J. Am. Soc. Nephrol.*, vol. 4, pp. 1441–1448, 2009.
- [4] J. C. Kim and C. Ronco, "Personal daily dialysis: the evolution of the artificial kidney," *Blood Purif. J.*, vol. 36, pp. 47–51, 2013.
- [5] J. C. Kim and C. Ronco, "The human nephron filter: toward a continuously functioning, implantable artificial nephron system," *Blood Purif. J.*, vol. 23, no. 4, pp. 269–274, 2005.
- [6] A. Davenport, "Portable and wearable dialysis devices for the treatment of patients with end-stage kidney failure: wishful thinking or just over the horizon?" *J. Pediatr. Nephrol.*, vol. 30, pp. 2053–2060, 2015.
- [7] A. R. Nissenson, "Bottom-up nanotechnology: the human nephron filter," *Semin. Dial.*, vol. 22, no. 6, pp. 661–664, 2009.
- [8] S. Takahashi, "Future home hemodialysis – advantages of the NxStage system one," *Contrib. Nephrol.*, vol. 177, pp. 117–126, 2012.
- [9] C. Ronco and L. Fecondini, "The Vicenza wearable artificial kidney for peritoneal dialysis (ViWAK PD)," *Blood Purif. J.*, vol. 25, pp. 383–388, 2007.
- [10] DM. Karantonis, MR. Narayanan, M. Mathie, NH. Lovell, BG. Celler, "Implementation of a Real-Time Human Movement Classifier Using a Triaxial Accelerometer for Ambulatory Monitoring" *IEEE Trans Inf Technol Biomed.*, vol. 10, no. 1, pp 156-167, 2006
- [11] MR. Narayanan, SR. Lord, MM. Budge, BG. Celler, NH. Lovell, "Falls management: detection and prevention, using a waist-mounted triaxial accelerometer", *Conf Proc IEEE Eng Med Biol Soc.* Pp 4037-4040, 2007
- [12] DW Kang, JS Choi, JW Lee, SC Chung, SJ Park, GR Tack, "Real-time elderly activity monitoring system based on a tri-axial accelerometer", *Disabil Rehabil Assist Technol.*, vol. 5, no. 4, pp 247-253, 2010

Data Mining Models Comparison for Diabetes Prediction

Amina Azrar¹, Muhammad
Awais³

Department of Software Engineering
Government College University,
Faisalabad, Pakistan

Yasir Ali²

Department of Computer Science
and Engineering
University of Engineering and
Technology,
Lahore, Pakistan

Khurram Zaheer⁴

Department of Software Engineering
Government College University,
Faisalabad, Pakistan

Abstract—From the past few years, data mining got a lot of attention for extracting information from large datasets to find patterns and to establish relationships to solve problems. Well known data mining algorithms include classification, association, Naïve Bayes, clustering and decision tree. In medical science field, these algorithms help to predict a disease at early stage for future diagnosis. Diabetes mellitus is the most growing disease that needs to be predicted at its early stage as it is lifelong disease and there is no cure for it. This research is intended to provide comparison for different data mining algorithms on PID dataset for early prediction of diabetes.

Keyword—Diabetes; data mining; classification; decision tree; Naïve Bayes; KNN

I. INTRODUCTION

Knowledge discovery in databases (KDD) is the system of applying data mining algorithms. Knowledge Discovery in Databases (KDD) is common research area for researchers in machine learning, databases, high performance computing, data visualization and knowledge-based systems. The primary steps for data mining include data selection, data preprocessing, data transformation, data mining, and final evaluation (pattern evaluation and pattern recognition).

Data Mining is the process of getting meaningful outcomes from any given dataset. Some of the techniques used for data mining include association rules, classification, clustering, Naïve Bayes, Decision Tree and KNN. A variety of rules can be generated using data mining techniques. Data Mining is useful for Prediction or Description of a few records. Using prediction, we are expecting unknown values of various variables in dataset whilst description specializes in coming across designs that depict the information translated by means of People.

Data mining is useful for predicting diseases. Affected person's history, Hospitals, clinical devices and electronic facts offer a lot of records concerning a selected disease. Those datasets are used for extracting useful information by which we are able to take choices and generate rules. Multiple diseases can be diagnosed using data mining methodologies, for example, AIDS and diabetes. This paper is meant to predict diabetes for pregnant women depending on few given attributes. Some major factors that affect the diabetes or may

cause its increase in severity include obesity, weight increase or hypertension.

A. Diabetes

Diabetes mellitus is a common disease where there is too much sugar (glucose) floating around in your blood. This occurs because either the pancreas can't produce enough insulin or the cells in your body have become resistant to insulin. Diabetes affects the capability of human body to utilize the energy present in food. Basic types of diabetes are:

Type1 – In this type of diabetes pancreas does not produce adequate amount of insulin and in consequence the level of glucose in blood exceeds from typical range. Individuals suffering from this type diabetes are usually dependent on external insulin injected in body after regular intervals. It is caused by a genetic predisposition. Medical risks associated with this type of diabetes include diabetic retinopathy (eyes disorder), diabetic neuropathy (nerves disorder) and diabetic nephropathy (kidneys disorder). It counts for 95% diabetes cases.

Type2 – In Type 2 diabetes body is unable to consume the insulin properly due to insulin resistance. It is usually caused due to obesity and overweight children. It is non-insulin dependent and milder than Type 1 diabetes. It causes major effects on heart diseases and heart strokes. It cannot be cured but controlled with proper nutrition, exercise and weight management.

Gestational Diabetes – This type of diabetes includes married women who are not affected with diabetes according to previous medical history but high glucose level is diagnosed during/after pregnancy. According to the National Institutes of Health, the reported rate of gestational diabetes is between 2% to 10% of pregnancies.

II. LITERATURE SURVEY

- Tawfik Saeed Zeki et al. [1] in their research presented an expert system for diabetes diagnosis. Their proposed expert system was rule based that have the structure of IF THEN. Transforming experts' knowledge to stated rules, they defined 3 stages that are handled by Block Diagram, Mockler Charts and Decision Tables. Total 6 states of diagnosis had been described (by the block diagram of diagnosis) using 5 attributes with different

combinations for various diagnosis. After inspecting multiple factors, expert system provides diagnosis for disease. It was coded in VP-Expert as it is specified for developing expert systems.

- Seyedeh Talayeh Tabibi et al. [2] proposed an expert system for checkup and treatment of different types of diabetes. Expert system was developed in 10 stages. They took 3 attributes that include patient's condition, patient's information and different tests. Multiple combinations are generated on the basis of given attributes as it was rule-based expert system. Questions are generated relating to test and background relating to diabetes and suitable advice is generated depending on situation. It was developed in VP-Shell to code while they also obtained experts advice from medical specialists and nurses of diabetics' department.
- Vishali Bhandari and Rajeev Kumar [3] compared Mamdani-type and Sugeno-type fuzzy expert systems with the help of multiple parameters for diabetes diagnosis. MATLAB fuzzy logic toolbox is used for comparative study for both types of expert systems. Different resulting parameters showed that Sugeno-type expert system is more useful as it less computational and optimized while Mamdani-type expert system is not computationally powerful. Mamdani-type fuzzy expert system generates outcome using defuzzification and has outcome membership functions while Sugeno-type uses weighted average for outcome and has no outcome membership function. 5 parameters are used and results are generated that are compared afterward.
- Neeru Lalka and Sushma Jain [4] presented an expert system for diagnosis and medication for Type-I diabetes. Multiple parameters are used that include body mass index (bmi), plasma glucose level, minimum blood pressure and serum insulin level. Specified dosage for insulin intake is recommended based on few attributes. Probability of diagnosis uses five fuzzy numbers to show results and also accurately calculates probability to avoid hypoglycemic (low level of blood sugar) condition. Three fuzzy numbers are used to show output results. This expert system is only for Type-I diabetes.

III. DATASET

A. Pima Indians Diabetes Data set

Dataset contains records of females, having age at-least 21 years and living in Phoenix, Arizona, USA. Dataset contains labeled data having class attribute in binary (0 or 1). 0 value of class attribute represents negative test and 1 value represents the diagnosis of diabetes. Dataset contains 768 records of different patients in which 268 (34.9%) records are positive test cases which means '1' value of class attribute and 500 (65.1%) cases in class '0' representing negative test. The attributes of dataset are given in Table I with their description, type and units.

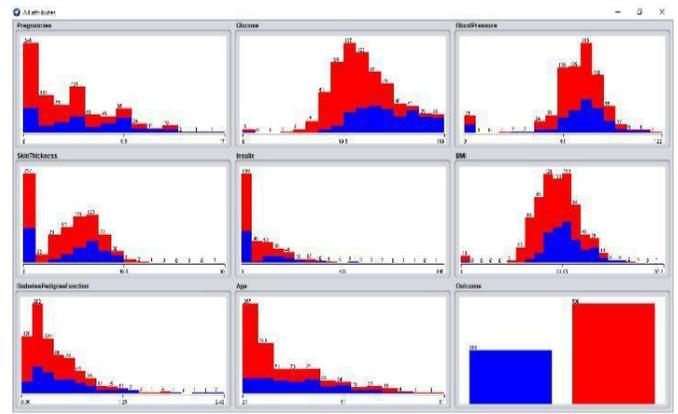


Fig. 1. Results of Class Label.

TABLE I. DATASET DESCRIPTION

Name	Description	Type	Unit
Pregnant	Number of pregnancies	Numeric	-
GTT	2-hour OGTT Plasma glucose	Numeric	mg/dl
Bp	Diastolic BP	Numeric	mmHg
Skin	Triceps Skin fold thickness	Numeric	Mm
Insulin	2-hour serum insulin	Numeric	Mm, U/ml
BMI	Body mass index (kg/m)	Numeric	Kg/m ²
DPF	Diabetes pedigree function	Numeric	-
Age	Age of Patient (years)	Numeric	-
Diabetes (Label)	Diabetes onset within 5 years (0,1)	Numeric	-

Visualization of dataset using WEKA presents the data distribution shown in figure. Fig. 1 shows 5 input variables and one outcome variable. Red color represents class label with negative results while blue color shows class label with values of labeled with negative result.

B. Preprocessing

Pre-processing consists of the steps of collection/cleaning, selection and transformation, data mining (integration and normalization) and last step is evaluation. Cleaning is used to fill the missing values in datasets using one-of-a-kind techniques like binning or replacing by mean or mode.

For pre-processing and applying few data mining algorithms, numerical data is converted to categorical data. Outcome is converted from integral data to categorical data with class labels as YES and NO while other categories are based on general items used and displayed in tables below (Tables II and III). Data for BP and Glucose is categorized on

the basis of general categories used and ranges defined in below table.

TABLE II. BP LEVEL RANGES [6]

BP Level	Range
Low	Less than 80
High	80 to 100
Hypertension	Above 100

TABLE III. GLUCOSE LEVEL RANGES

Glucose Label	Range
Low	Less than 80
Normal	80 to 140
Early Diabetes	141 to 180
Diabetes	Above 181

IV. APPLIED ALGORITHMS

1) *Decision Tree*: Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It is tree like a graph used to display every possible outcome of a decision. It is most powerful classification algorithm used to predict possible outcome of a branch or tree. Classification is done by tree and leave nodes are generated on the basis of results on nodes in it.

Parameters on the dataset when applying DT set as criterion was gain ratio, maximal depth of the tree considered as 20. We also applied pruning as confidence=0.25 and pre-pruning techniques on DT as by setting minimal gain=0.1, minimal leaf size=2, minimal size of split=2 and number of pruning alternates considered as 3 in both datasets. We split data in DT as 70% training data and 30% as test data and apply model to show outcome and performance to check effectiveness and accuracy of both treatments.

Result using information gain show the class precision of yes and no and the Accuracy of decision tree algorithm up to 75.65%.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig. 2. Probability Calculation [5].

Results using Gain Ratio

	Yes	No
Yes	44	20
No	36	130

2) *Naïve Bayes*: Naïve Bayesian is a well-known type of data mining classification technique. According to the definition it is a statistical technique which predicts the class of a new data record by estimation governed by the probabilities calculated from Bayesian rule formula. The Naïve Bayesian basic principle can be described as: Calculation of probability of Hypothesis that record belongs to class c given the new observed data record x.

Training process includes the calculation of marginal and conditional probabilities which are used in testing process for the calculation of probability of belonging of a new record to any class (Fig. 2).

Result using Naïve Bayes

	Yes	No
Yes	46	31
No	34	119

Result using algorithm show the class precision of yes and no, the Accuracy of Naïve Bayes algorithm shows 71.74% and Distribution model for label attribute Outcome is:

- Class **Yes (0.349)** -> 8 distributions
- Class **No (0.651)** -> 8 distributions

3) *K-Nearest Neighbor (KNN)*: K-Nearest Neighbor (KNN) is supervised learning algorithm used for classification of data. K means to select points from given dataset that how much data will be selected of nearest neighbor. This algorithm selects data on the basis of K value to nearest neighbor and decides that this point is similar to given sample. We apply KNN on dataset with K values ranging from 1 to 10. First, we make label to results of the treatment and split data into 70%, 30% as training and test records respectively, and then we make 10-fold of cross validation, also with 20 folds, by giving sampling as automatic to the split data value and apply KNN on the given data.

For 10 Folds (Fig. 3):

Neighbors (K)	Accuracy
1	64.84%
2	60.20%
3	64.28%
4	62.61%
5	63.92%

Fig. 3. Accuracy using 10-Fold.

For 20 Folds (Fig. 4):

Neighbors (K)	Accuracy
1	65.19%
2	60.58%
3	65.04%
4	63.36%
5	65.05%

Fig. 4. Accuracy using 20-Fold.

By applying K-nearest Neighbor algorithm, we find out several accuracies using 10 and 20-fold in KNN algorithm using 1 to 5 nearest neighbors on dataset. Using 1 nearest neighbor in 20-fold, highest accuracy can be seen.

V. RESULTS

The results obtained from these 3 applied algorithms are different that as each algorithm worked on different technique. Results obtained from this dataset can be enhanced by applying more pre-processing techniques and data filtration. Accuracy obtained from Decision Tree is highest yet the graph is more dispersed that can be enhanced too. Lowest accuracy is from KNN. KNN is tested with wide range of K values from 1 to 10 and with changing folds from 10 to 20 but still accuracy is not that much. Pictorial representation of results is shown below in the form of graph.

Comparisons

Fig. 5 shows the accuracy rate of different data mining (DM) models.

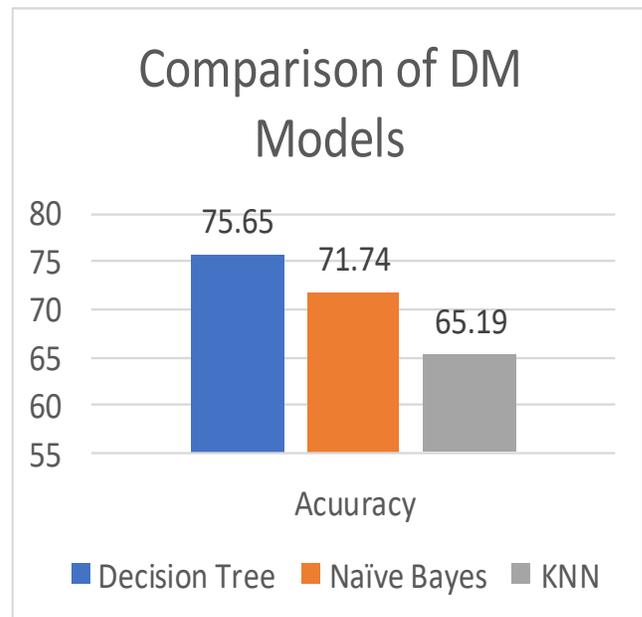


Fig. 5. Comparison of Different DM Models.

VI. CONCLUSIONS

The prevalence of diabetes is increasing among young adults and old age people. This paper focuses that the use of data mining algorithms can be very helpful in early prediction and in consequence early precautions before the diagnosis of disease. The main goal of this paper is to provide a comparison and suggest best algorithm which can be used for the pattern recognition or prediction in healthcare fields. These algorithms are of much importance for medical datasets because these algorithms can be used for automatic classification tools which can help doctors or experts for taking necessary steps for any disease before diagnosis. Each of these algorithms can give high accuracy and efficiency depending upon the type of data and attributes. After the implementations of these algorithms it can be said that for PID dataset Decision Tree gives best accuracy 75.65%. The tool used for testing and validation is Rapid Miner while all algorithms worked with 70:30 ratio for training and testing.

REFERENCES

- [1] Tawfik Saeed Zekia, Mohammad V. Malakootib, Yousef Ataepoor, S. Talayeh Tabibid. An Expert System for Diabetes Diagnosis. American Academic & Scholarly Research Journal Special Issue Vol. 4, No. 5, Sept 2012.
- [2] Seyedeh Talayeh Tabibi, Tawfik Saeed Zaki, Yousef Ataepoor. Developing an Expert System for Diabetics Treatment Advices. International Journal of Hospital Research 2013, 2(3):155-162.
- [3] Vishali Bhandari and Rajeev Kumar. Comparative Analysis of Fuzzy Expert Systems for Diabetic Diagnosis. International Journal of Computer Applications (0975 – 8887) Volume 132 – No.6, December 2015.
- [4] Neeru Lalka and Sushma Jain. Fuzzy Based Expert System for Diabetes Diagnosis and Insulin Dosage Control. International Conference on Computing, Communication and Automation (ICCCA2015).
- [5] (Data Mining Map, an Introduction to Data Science, 2010-2018)
- [6] (Blood Pressure UK, 2008).

Using Artificial Intelligence Approaches to Categorise Lecture Notes

Naushine Bibi Baijoo, Khusboo
Bharossa
Faculty of Engineering
University of Mauritius
Reduit, Mauritius

Somveer Kishnah
Software and Information Systems
Dept
University of Mauritius
Reduit, Mauritius

Sameerchand Pudaruth
ICT Department
University of Mauritius
Reduit, Mauritius

Abstract—Lecture materials cover a broad variety of documents ranging from e-books, lecture notes, handouts, research papers and lab reports amongst others. Downloaded from the Internet, these documents generally go in the Downloads folder or other folders specified by the students. Over a certain period of time, the folders become so messy that it becomes quite difficult to find our way through them. Sometimes files downloaded from the Internet are saved without the certainty that they will be used or revert to in the future. Documents are scattered all over the computer system, making it very troublesome and time consuming for the user to search for a particular file. Another issue that adds up to the difficulty is the improper naming conventions. Certain files bear names that are totally irrelevant to their contents. Therefore, the user has to open these documents one by one and go through them to know what the files are about. One solution to this problem is a file classifier. In this paper, a file classifier will be used to organise the lecture materials into eight different categories, thus easing the tasks of the students and helping them to organise the files and folders on their workstations. Modules each containing about 25 files were used in this study. Two machine learning techniques were used, namely, decision trees and support vector machines. For most categories, it was found that decision trees outperformed SVM.

Keywords—Classification; lecture materials; machine learning; support vector machines; decision trees

I. INTRODUCTION

The rapid advancements in IT have brought about an exponential increase in the number of electronic documents. Documents that were presented on paper in the past are today created, stored, distributed and displayed digitally [1]. This trend has captured a wide variety of fields, if not all. The education field has not been left behind in the process. It has evolved alongside with the advent of new technologies.

Students nowadays have thousands of files on their workstations, scattered in different folders, on different drives, etc. Some files have meaningful names while others do not. The easy access to information has also led to an increase in the amount of irrelevant information. Information from web pages, news articles, presentations, papers are saved on the machines without the certainty that they will be of some use in the future. This usually costs users a great deal of time looking for a particular file especially if all the files are scattered in different places on the computer system and the file in question

is not properly named. Therefore, an automatic file classification system is of utmost importance. The role of the file classifier would be to go through all the files in a given folder and determine the best fitting category for each file.

This paper proceeds as follows. Section II gives a description of the different techniques that are used for the classification process. Section III describes the methodology used and the tasks that need to be carried out to classify the documents. Section IV outlines the implementation process and critically analyses and evaluates the results of the classifiers. Finally, we conclude the study in Section V.

II. LITERATURE REVIEW

A. Text Mining

Text mining, also known as text analytics, is a hypernym used to describe the wide range of technologies in place to analyze and process unstructured and semi-structured textual data [2], [3]. These technologies are used to extract meaningful information from documents or files that would then serve particular purposes. The most common theme behind all the technologies is to turn textual information into numbers. Algorithms are then applied to the numerical format of the words, documents and eventually to full databases. The data is then handled and processed as per to one's requirements.

Text mining involves the applications of techniques from fields such as information retrieval, information extraction, natural language processing, machine learning, classification, clustering and text categorisation. Information retrieval is an area pertaining to the organisation, examination, storage and retrieval of information from different sources. It performs several tasks such as document ranking and document classification. This paper discusses two main classification techniques, namely decision trees and support vector machines.

B. Decision Trees

Decision trees are a very simple but powerful classification method. One advantage of a decision tree is that it can be very easily interpreted by humans. It is commonly used in pattern recognition problems for knowledge systems [4]. A decision tree is very similar to a flow diagram. It consists of an internal node with many attached branches and leaf nodes. A test on a particular element is designated by the internal node. The branches denote the result of that experiment and finally, the class distribution is indicated by the leaf nodes [5]. The

topmost node is known as the root node and it is denoted by an oval. Rectangles are used to symbolise the internal nodes. The leaf nodes, on the other hand, are circular in shape.

A list of attributes is made for measurement in order to create a decision tree. A target attribute is then chosen for prediction. All data is processed to know the number of times an attribute appears in each document. Decision trees use the concept of entropy for splitting attributes – reducing the number of attributes. Splitting the attributes results in a hierarchy of branches. These branches or nodes are called the decision tree. All nodes can form another branch of node. Each branch in the tree produces an observation. This observation is made using the state of one of the fields in the dataset. Another method used for splitting is called pruning. There are two types of well-known pruning namely pre-pruning and post-pruning also known as forward pruning and back-pruning respectively. In pre-pruning, the user decides when to stop adding attributes during the building process. As a result, it can lead to very biased decisions as individual attributes do not contribute much to the decision. Post-pruning is different in that the decision tree is fully built prior to pruning the elements [6].

Decision trees are efficient for new and unseen inspections. However, building a decision tree can be very time-consuming. One serious weakness of decision trees is the problem of error propagation throughout a tree. Decision trees are built by a series of local decisions. These local decisions have a carry-over effect. Therefore, if one of the local decisions goes wrong at some point in time, all successive decisions are bound to be bad as well. In such a case, the correct path of the tree might not be returned [6].

C. Support Vector Machines

SVM algorithms are a learning method introduced by Vladimir Vapnik and colleagues. They are used for pattern recognition, classification and regression. Support vector machines have been very successful in various learning areas [7], [8]. SVMs construct hyperplanes for linearly separated patterns. The basic idea in SVM is to find a mediator which separates multi-dimensional data into two classes [9]. SVMs work towards maximising predictive accuracy while avoiding over-fitting. SVMs give very significant results for applications involved in classifying text, recognizing hand-written characters, classifying images and also in bio-informatics. One of the strongest points for SVMs is that they impose no limit on the number of attributes that can be used. However, the only problem is that SVMs require a lot of memory [10].

III. METHODOLOGY

The very first step to the classification of the lecture materials is to build a dataset. A dataset in this study is simply a bulk of relevant documents. Eight categories of lecture materials amounting to 213 files were selected and were put in a common folder. Table I shows the categories and the number of files used in each category.

NLTK (Natural Language Toolkit) has been used to process the files. It is the most commonly used platform to write Python programs to interact with textual data [11]. It is open source software and is made up of a plethora of libraries to allow for the manipulation of high-level data.

TABLE I. DETAILS OF CATEGORIES

Category	Number of files
Cyberlaws	23
Database	35
Enterprise Resource Planning	25
Management Information Systems	26
Multimedia	26
Networking	33
Security	24
Software Engineering	21

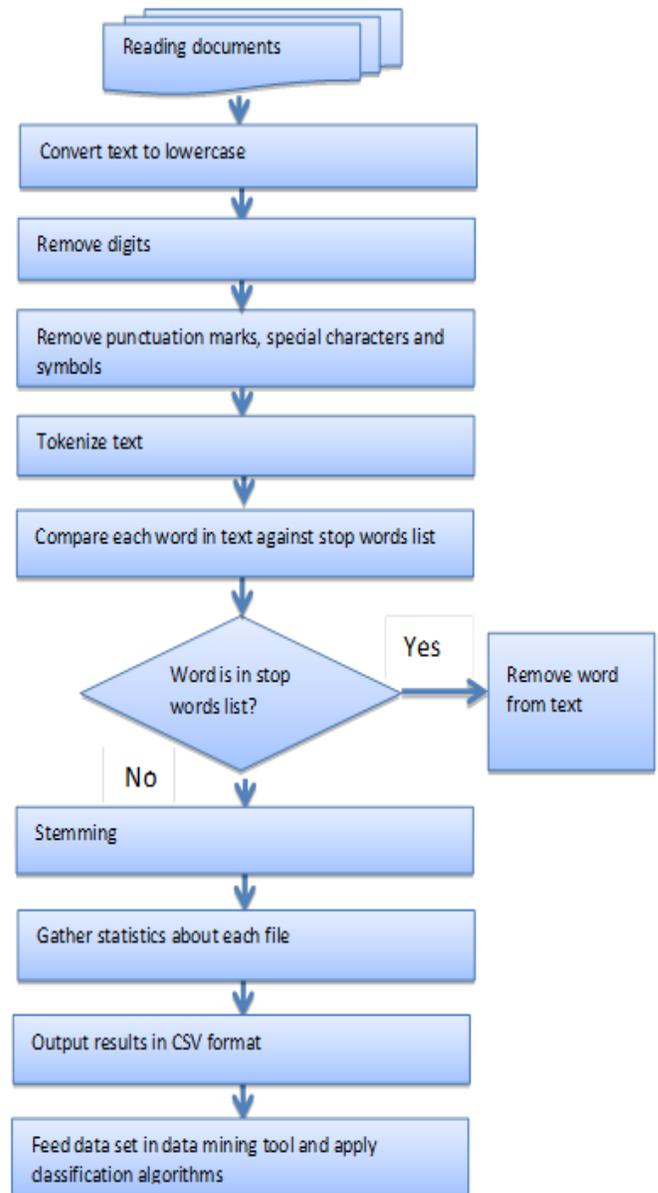


Fig. 1. Flowchart Outlining the Steps of the Implementation Process.

Firstly, the documents are converted to lowercase to avoid ambiguities at later stages. Secondly, the files are cleaned. All the punctuation marks, special symbols, digits and special characters are removed. The series of words is then subjected to the process of tokenization which breaks the documents into distinct words or tokens. Each word is then checked against NLTK's stopword list. The stopword list is a large body of text consisting of 11 languages with a total of 2,400 stop words [12]. Stop words are words like 'the', 'is', 'a', that do not carry much weight when it comes to determining the best category of a file. Thus, all stop words are eliminated from the documents leaving us with only potentially useful and meaningful words.

The last step in the cleaning process is the application of stemming to the words, as shown in Fig. 1. Stemming is a method for removing the affixes from a word in order to end up only with the stem which is also known as the root. It is a common technique used in search engines for indexing words. The search engine stores only the stems, instead of keeping all the different forms of a word. This is very helpful as it reduces the size of the index by a considerable amount, thus improving performance and retrieval accuracy. One of the most popular stemming algorithms is the Porter Stemmer Algorithm. It removes and replaces well known suffixes of English words [13]. NLTK supports a number of other stemming algorithms as well, namely the Lancaster stemmer, Regexp stemmer and the Snowball stemmer [14]. For this project, the Snowball stemmer has been used.

Once the documents are cleansed, the array of meaningful and stemmed words is further processed to get the frequencies of each word in each document. The outputs are stored in CSV files. These CSV files produced are fed into WEKA [15]. The following section gives more details about the classification process in WEKA and evaluates the classifier outputs.

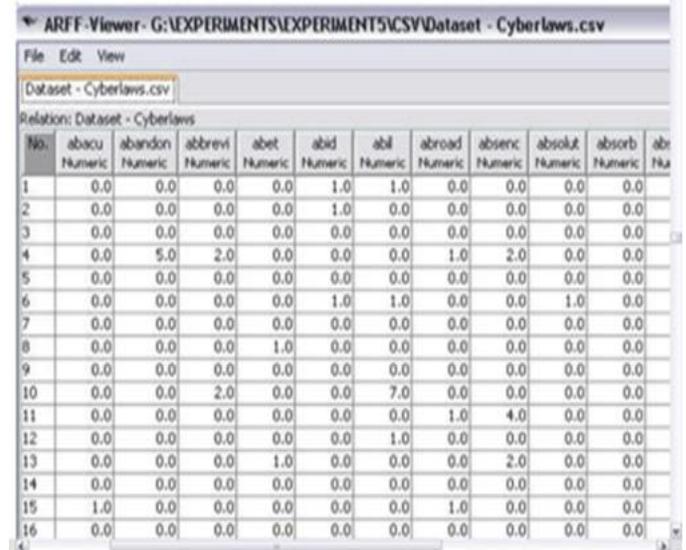
IV. IMPLEMENTATION AND EVALUATION

WEKA supports a particular file format known as the ARFF data format. ARFF stands for Attribute – Relation File Format. It is an ASCII file describing a set of samples having a number of elements in common. The ARFF-Viewer tool in WEKA allows for the conversion of CSV data files to the ARFF data format. An ARFF data file has a very particular format. It basically has two distinct sections, the header part followed by the data information. It starts with @RELATION, which gives the name of the file, followed by @ATTRIBUTE, giving a list of the file's attributes and lastly @DATA.

All the attributes in an ARFF file are of type 'numeric' since we are dealing with the frequencies of the words in the documents. The data is represented as a stream of numbers. Viewed in WEKA's ARFF-Viewer, we are presented with a tabular form of the file (Fig. 2), which is easier to interpret.

The datasets for all eight categories of lecture materials were classified using two different machine learning techniques and the outputs were compared. From existing works, we have noticed that it is a common practice to test the algorithms with a balanced number of positive and negative samples. Thus, we have used an equal number of documents to carry out the experiments. A binary approach was followed, i.e.

for each category we took 15 positive samples and 15 negative samples (which was termed as the 'Others' category).



No.	abacu Numeric	abandon Numeric	abbrevi Numeric	abet Numeric	abid Numeric	abil Numeric	abroad Numeric	absenc Numeric	absolut Numeric	absorb Numeric	ab N/A
1	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	5.0	2.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
6	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
8	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	0.0	0.0	2.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	
11	0.0	0.0	0.0	0.0	0.0	0.0	1.0	4.0	0.0	0.0	
12	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
13	0.0	0.0	0.0	1.0	0.0	0.0	0.0	2.0	0.0	0.0	
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
15	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Fig. 2. The ARFF-Viewer.

A. J48

The datasets were first classified using the J48 decision tree algorithm in WEKA. J48 normally selects a set of keywords in the set to base its decision on [16]. However, the selection of that keyword is not stable as a little change in the dataset may alter the results by a great amount. Also, the keyword chosen may not always reflect the intended category. An example is given in Fig. 3.

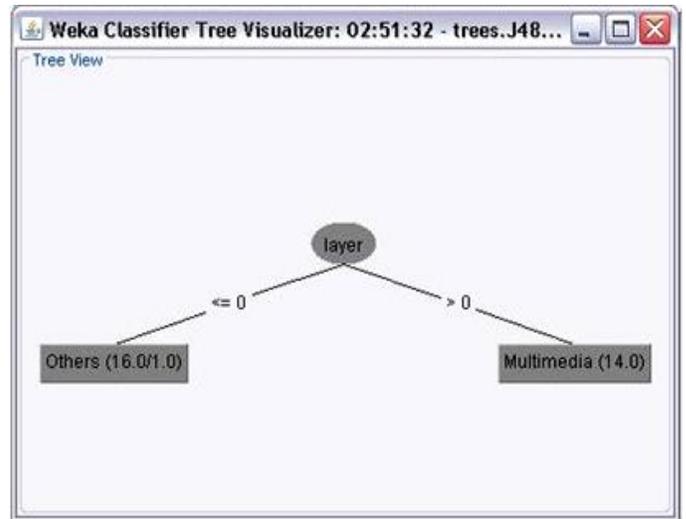


Fig. 3. Decision Tree for Multimedia.

Fig. 3 shows the classifier's tree visualizer for Multimedia. The word 'layer' has been chosen to decide between the Multimedia and the Others categories. This word however is not appropriate as it may be used in many contexts other than Multimedia. Words like 'multimedia', 'image', 'video' would have been more appropriate in this case.

B. LibSVM

The datasets were subjected to a second round of classification, this time with LibSVM [17]. The classification for the Multimedia category, for instance, yielded very good results. All of the 15 documents pertaining to this category were correctly classified.

TABLE II. CONFUSION MATRIX

Predicted Class			Actual Class
	Others	Multimedia	
Others	7	0	
Multimedia	8	15	

Table II indicates that out of 15 files that are actually from the Others category, seven of them were correctly classified while the remaining eight were not. They were classified as Multimedia files instead of Others. As for the Multimedia files, it was an error-free classification.

C. Summary of Outputs

Table III shows a summary of the classifier outputs with J48 and LibSVM for all the 8 categories of lecture materials.

A pertinent observation is the meagre percentage of correctly classification instances for the Database category. Database is a very common field in computing. It merges with many other fields in a fluid manner and it may be applied in a variety of computing contexts. Therefore, files from Enterprise Resource Planning (ERP) and Management Information Systems (MIS) files may well fall in the Database category. This is one potential reason for the downfall in the positive percentage for this particular category.

TABLE III. SUMMARY OF CLASSIFIERS OUTPUTS

Categories	Correctly classified instances	Incorrectly classified instances	Correctly classified instances	Incorrectly classified instances
	J48		LibSVM	
Cyberlaws	23	7	25	5
	76.7%	23.3%	83.3%	16.7%
Database	19	11	20	10
	63.3%	36.7%	66.7%	33.3%
Enterprise Resource Planning	24	6	22	8
	80%	20%	73.3%	26.7%
Management Information Systems	24	6	22	8
	80%	20%	73.3%	26.7%
Multimedia	26	4	22	8
	86.7%	13.3%	73.3%	26.7%
Networking	27	3	25	5
	90%	10%	83.3%	16.7%
Security	28	2	26	4
	93.3%	6.7%	86.7%	13.3%
Software Engineering	29	1	22	8
	96.7%	3.3%	73.3%	26.7%

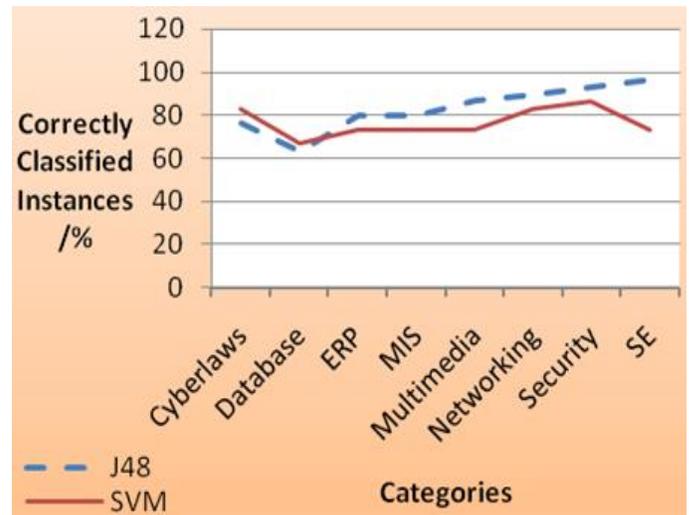


Fig. 4. Line Graph Comparing Results of J48 with SVM.

The overall accuracy for J48 is 83.3% while for SVM it was 76.7%. From these statistics and from Fig. 4, we can see that J48 has done slightly better in this scenario.

D. Accuracy of Outputs

The accuracy of the classifier outputs in WEKA is determined by some very distinct parameters. These parameters are: True Positive Rate (TP Rate or Recall), False Positive Rate (FP Rate), Precision and the F-measure.

TABLE IV. ACCURACY BY CATEGORY

Categories		TP Rate	FP Rate	Precision	F-measure
Cyberlaws	J48	0.667	0.133	0.8	0.741
	LibSVM	0.933	0.267	0.778	0.848
Database	J48	0.733	0.467	0.611	0.667
	LibSVM	0.4	0.067	0.857	0.545
ERP	J48	0.867	0.267	0.765	0.813
	LibSVM	0.6	0.133	0.818	0.692
MIS	J48	0.8	0.2	0.8	0.800
	LibSVM	0.933	0.467	0.667	0.778
Multimedia	J48	0.933	0.2	0.824	0.875
	LibSVM	0.467	0	1	0.636
Networking	J48	0.933	0.133	0.875	0.903
	LibSVM	0.8	0.133	0.857	0.828
Security	J48	1	0.133	0.882	0.938
	LibSVM	1	0.267	0.789	0.882
Software Engineering	J48	1	0.067	0.938	0.968
	LibSVM	0.667	0.2	0.769	0.714

Table IV shows the accuracy by category for both classifiers. A TP rate of one is an ideal result. It means that all or almost of the documents were correctly classified. All Security files were correctly classified, hence yielding a recall of 100% with both classifiers. Fields like Software Engineering and Cyberlaws, which are quite distinct from the rest, have also fetched high values. The recall value for Multimedia is exceptionally low for the SVM classifier. However, the explanation for this can be seen in Table II. This is because many files from the Others category were classified as being in the Multimedia category due to the presence of certain superfluous words. Nevertheless, the precision values are very high. A TP rate as low as 0.4 is an undesirable result, which is indicative of poor classification of the files. It is noticed that the TP rates for ERP and MIS are not very high too. These values point towards the confirmation of the observation that the modules ERP, MIS and Database bear a lot of similar words, hence some files were incorrectly classified. In general, the values for precision and recall were appreciably high.

V. CONCLUSIONS

This paper discussed the classification of lecture materials. Two hundred and thirteen documents from eight different university modules were selected and were classified into pre-defined sets. The documents were classified using two different machine learning techniques namely decision trees and support vector machines. A number of experiments were carried out and the results of the classification were critically analysed. The outputs' parameters and various other factors showed that J48 was a better classification technique than SVM for this particular case. The overall accuracy for J48 was found to be 83.3% while for SVM it was only at 76.7%. However, these results cannot be generalised as our data set was quite small. In the future, we intend to repeat these experiments with many more files and more classifiers such as kNN, Naïve Bayes and artificial neural networks. Document size, i.e. the number of words in each file will also be taken into consideration.

REFERENCES

- [1] H. Abdulkadhim, M. Bahari, A. Bakri, and W. Ismail, "A research framework of Electronic Document Management Systems (EDMS) implementation process in Government," *Journal of Theoretical and Applied Information Technology*, vol. 81(3), pp. 420-432, 2015.
- [2] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Oxford: Academic Press, 2012.
- [3] S. Binkheder, H. Y. Wu, S. Quinney, and L. Li, "Analyzing Patterns of Literature -Based Phenotyping Definitions for Text Mining Applications," 2018 IEEE International Conference on Healthcare Informatics (ICHI), 4-7 June 2018, New York, USA.
- [4] Y. Ben-Haim, and E. Tom-Tov, "A Streaming Parallel Decision Tree Algorithm," *The Journal of Machine Learning Research*, vol. 11, pp. 849-872, 2010.
- [5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concept and Techniques*, 3rd ed. MA: Morgan Kaufmann, 2011.
- [6] M. Jaworski, P. Duda, and L. Rutkowski, "New Splitting Criteria for Decision Trees in Stationary Data Streams," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29(6), pp. 2516-2529, 2018.
- [7] S. Tong, and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45-66, 2002.
- [8] Y. You, J. Demmel, K. Czechowski, L. Song, and R. Vuduc, "Design and Implementation of a Communication -Optimal Classifier for Distributed Kernel Support Vector Machines," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28(4), pp. 974-988, 2017.
- [9] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," 10th European Conference on Machine Learning (ECML), pp. 137-142, 21-23 April 1998, Chemnitz, Germany.
- [10] B. Bryant, H. Sari-Sarraf, R. Long, and S. Antani, "A Kernel Support Vector Machine Trained Using Approximate Global and Exhaustive Local Sampling," 4th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, pp. 267-268, 5-8 December 2017, Texas, USA.
- [11] NLTK 3.3 Documentation. Retrieved January 4, 2018, from: <http://www.nltk.org/>
- [12] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol: O'Reilly Media, 2009.
- [13] J. Perkins, *Python Text Processing with NLTK 2.0 Cookbook*. Birmingham, Packt Publishing Ltd, 2010.
- [14] A. Schofield, and D. Mimno, "Comparing Apples to Apple: The Effects of Stemmers on Topic Models," *Transactions of the Association for Computational Linguistics*, vol. 4, 287-300, 2016.
- [15] Weka 3: Data Mining Software in Java. Retrieved January 10, 2018, from: <http://www.cs.waikato.ac.nz/ml/weka/>
- [16] F. Borges, R. Fernandes, A. M. Lucas, and I. Silva, "Comparison Between Random Forest Algorithm and J48 Decision Trees Applied to the Classification of Power Quality Disturbances", 11th International Conference on Data Mining (DMIN), 27-30 July 2015, Las Vegas, Nevada.
- [17] F. Wang, Y. Yang, C. Liu, H. Wang, and X. Weng, "Analysis of Influencing Factors of Water Traffic Accidents based on LIBSVM", 27th International Ocean and Polar Engineering Conference, 25-30 June 2017, San Francisco, California.

EEG-Based Emotion Recognition using 3D Convolutional Neural Networks

Elham S.Salama, Reda A.El-Khoribi, Mahmoud E.Shoman, Mohamed A.Wahby Shalaby

Information Technology Department
Faculty of Computers and Information, Cairo University
Cairo, Egypt

Abstract—Emotion recognition is a crucial problem in Human-Computer Interaction (HCI). Various techniques were applied to enhance the robustness of the emotion recognition systems using electroencephalogram (EEG) signals especially the problem of spatiotemporal features learning. In this paper, a novel EEG-based emotion recognition approach is proposed. In this approach, the use of the 3-Dimensional Convolutional Neural Networks (3D-CNN) is investigated using a multi-channel EEG data for emotion recognition. A data augmentation phase is developed to enhance the performance of the proposed 3D-CNN approach. And, a 3D data representation is formulated from the multi-channel EEG signals, which is used as data input for the proposed 3D-CNN model. Extensive experimental works are conducted using the DEAP (Dataset of Emotion Analysis using the EEG and Physiological and Video Signals) data. It is found that the proposed method is able to achieve recognition accuracies 87.44% and 88.49% for valence and arousal classes respectively, which is outperforming the state of the art methods.

Keywords—*Electroencephalogram; emotion recognition; deep learning; 3D convolutional neural networks; data augmentation; single-label classification; multi-label classification*

I. INTRODUCTION

Human emotions are important in communication with others and decision making. Recognizing emotion is important in intelligent Human-Computer Interaction (HCI) applications such as virtual reality, video games, and educational systems. In the medical domain, the detected emotions of patients could be used as an indicator of certain functional disorders, such as major depression. Human emotions are extracted from the facial expressions as the main source of emotions [1]. However, it is known that some people could hide their real emotions using misleading facial expressions [2]. Hence, researchers adhere to use other sources of information that are reliable and not susceptible to fraud. One of these sources is the electroencephalogram (EEG) signals which are the recording of the electric field of the human brain. The EEG signals are able to be used as a source of emotions since human responses are linked to the cortical activities. Ekman [3] found that emotion recognition needs to work under keeping expression in long duration. In other words, emotion-related signals contain contextual temporal dependencies. Hence, taking into consideration the relation in time between the EEG signal segments can model the bundling behavior of human emotions. In addition, the spatial relationship between multiple electrodes positions can prove that the behaviors of

human emotions are not isolating. However, most existing emotion recognition methods based on the EEG signals model only either spatial or temporal dependency. In this paper, a new emotion recognition method is proposed to extract spatiotemporal features from the EEG signals in one end-to-end model.

The main contributions of the proposed work are summarized as follows:

- The proposed work introduces a new approach which utilizes the 3D-CNN for extracting the spatiotemporal features in the EEG signals. To the best of our knowledge, employing the 3D-CNN has not yet been investigated for the EEG-based emotion recognition.
- The 3D-CNN captures the correlation between different channels positions by taking the data from different channels as input.
- The 3D-CNN proves its ability to capture the correlation between dimensional emotions (i.e. valence and arousal). This ability helps in converting the dimensional emotions into discrete emotions (i.e. happy, sad, etc.) and to save processing time needed for processing each dimensional label separately.
- The proposed 3D-CNN for the EEG-based emotion recognition has a significant potential to detect emotions from spatiotemporal information.

The rest of this paper is organized as follows: The previous and the most related research works are presented in Section II. Section III explains the proposed approach in details. The results are shown in Section IV. The proposed work is concluded in Section V.

II. RELATED WORKS

In well-documented works, the ability of the EEG signals for recognizing emotions was extensively explored [4], [5]. Verma and Tiwary [6] reported the use of the EEG signals for emotion recognition using the power spectral density as features and the Support Vector Machines (SVM) and the k-Nearest Neighbors (KNN) as classifiers. Yoon and Chung [7] introduced a new emotion recognition method using the EEG signals. They extracted features using the Fast Fourier Transform (FFT) analysis from the EEG segments and used the Pearson correlation coefficient for feature selection. A probabilistic classifier based on Baye's theorem is proposed.

In addition, a supervised learning using a perceptron convergence algorithm is introduced.

Naser and Saha [8] used the Dual-Tree Complex Wavelet Packet Transform (DT-CWPT) to extract meaningful emotion features from the EEG signal elicited during watching music videos. For feature elimination, Singular Value Decomposition (SVD), QR factorization with column pivoting (QRcp), and F-ratio are employed. The classification step is performed using SVM. Atkinson and Campos [9] introduced a novel feature-based emotion recognition model in which statistical features were extracted from the EEG segments such as median, standard deviation, and kurtosis coefficient. In addition to statistical features, band power (BP) for different frequencies, Hjorth parameters (HP), and fractal dimension (FD) for each channel are also extracted. This new model combined mutual information from feature selection methods and kernel classifiers. In order to reduce information redundancy, the minimum-Redundancy-Maximum-Relevance (mRMR) is employed. This method obtained the best set of features by selecting the features that are mutually different and have a high correlation. SVM classified the input data into low/high valence or low/high arousal.

Li et al. [10] proposed a new deep learning hierarchy of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to extract spatiotemporal features for emotion recognition from the EEG signals. The CNN is used for extracting the spatial features and its output is used as inputs to the RNN to extract the temporal features. In addition, Chen et al. [11] used four physiological signals including the EEG signals for emotion recognition using Hidden Markov Model (HMM) as a classifier. For feature selection, they utilized multimodal feature sets and Davies-Bouldin index (DBI) methods.

Koelstra et al. [12] extracted 216 EEG features from 5 different frequency bands. These features are theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30+ Hz), spectral power for 32 electrodes, and the difference between the spectral powers of all the symmetrical pairs of electrodes. For feature elimination, Fisher's linear discriminant was used and the Gaussian naive Baye's is used for the classification. Rozgic et al. [13] addressed emotion recognition based on the EEG signals and three classifiers: Neural Network (NN), NN voting, and SVM. They extracted the same extracted features in Koelstra et al. [12] from the EEG signals.

Alhagry et al. [14] extracted temporal features using RNN and the EEG signals. Their RNN consists of fully connected two LSTM layers, a dropout layer, and a dense layer. Zhang et al. [15] presented a deep learning framework called spatiotemporal recurrent neural network (STRNN) in order to combine the learning of spatiotemporal features for emotion recognition using the SJTU Emotion EEG Dataset (SEED).

However the accuracies obtained by the above researches are reasonably high, further improvement concerning emotion recognition is still needed.

III. PROPOSED SYSTEM

Normally, the automatic emotion recognition process can be carried out using one or more of different modalities: face, speech, body gestures, and the EEG signals. Using the EEG signals, researches focus on solving the problem of correlation in time between emotions. By nature, emotions last for short or long period of time, not just a moment. Thus, the relation between emotion segments in time is highly effective for improving recognition accuracy. Motivated by the recent success of the deep learning approaches [16], [17], the 3D-CNN approach is proposed to model the spatiotemporal information from the EEG signals. To reach this objective, data augmentation phase is first applied to increase the number of available EEG samples. Then, the 3D representation of inputs is created from the EEG segments. And finally, the proposed system of the 3D-CNN model is built. The procedure of the proposed system is illustrated in Fig. 1.

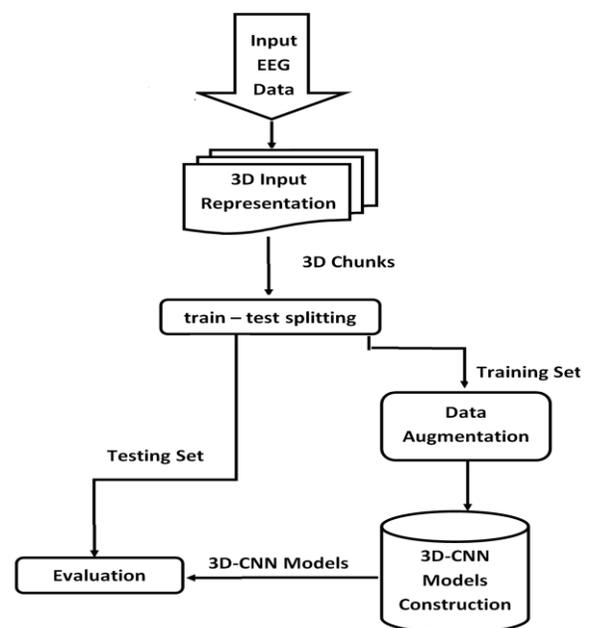


Fig. 1. The Flowchart of the Proposed System.

A. Data Augmentation

To evaluate the proposed system, the DEAP (Dataset of Emotion Analysis using EEG and Physiological and Video Signals) [12] data is used. It is a benchmark dataset for emotion analysis using the EEG, physiological and video signals. Thirty-two participants were watching 40 videos each with one-minute duration. The facial expressions and the EEG signals were recorded for each participant. The EEG signals were recorded from 32 different channels. Most of the publicly available EEG datasets have fewer amounts of data per participant. For the DEAP data, there are a limited number of samples; only 40 experiments are recorded per participant which may affect the performance of any machine learning system to generalize unseen samples. Data augmentation aims to increase the number of samples by adding some noise signals to the original input signals to generate new noisy samples and then train the model with these new noisy

samples [10]. This helps parameters to converge, avoid over-fitting, and make the proposed system capable of generalization to unseen samples.

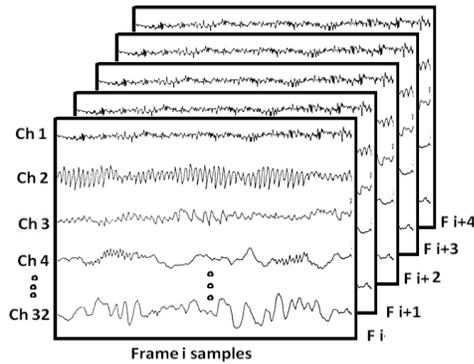


Fig. 2. The Representation of the 3D-CNN Input Volume with 5 Consecutive Frames.

To generate the noisy EEG signals, a Gaussian noise signal n with zero mean and unit variance is first generated randomly with N samples such that N is the number of samples of the original EEG signal s . Finally, the noisy version ξ of s is obtained by adding all samples of s and ξ signals together. The augmentation phase is applied in the training step only. During the testing step, the clean versions of the signals are used.

B. 3D Input Representation

As mentioned earlier, the 3D-CNN is capable of learning spatiotemporal features. This requires a construction of 3D input representations from the EEG signals. To this end, a 3D representation procedure is presented in the proposed work. Usually, the EEG data from every signal is recorded from different Ch channels. Using a window size w , the data from every channel c is segmented into small segments (frames). The number of frames from every channel is D frames (f_1, f_2, \dots, f_D). The samples of the i^{th} frame from all Ch channels are appended together to form a 2D matrix K where its height is the number of channels and its width is the number of samples in the i^{th} frame. Then, the third temporal domain is appended by selecting a number of consecutive frames m which is also called the chunk size. If the chunk size is 6, then, 6 sequential frames are appended together in one chunk in a 3D matrix called B .

To add a label to each chunk, the majority rule is employed to get the corresponding ground truth label. If the chunk has 6 frames and the same label occurs in more than 3 frames (chunk size / 2), this majority label is assigned to this chunk. Finally, a new 3D matrix C is created to hold the chunk of frames and its corresponding label. Each 3D matrix C is considered as an input to the 3D-CNN model for the training. Fig. 2 shows the shape of the 3D input volume. The figure shows a chunk with 5 consecutive frames and 32 channels.

C. The 3D-Convolutional Neural Network Model

The next step is to use the proposed the 3D-CNN to recognize emotions based on spatiotemporal features. The 3D-CNN structure is introduced and the proposed network architecture is thus described in the next subsections.

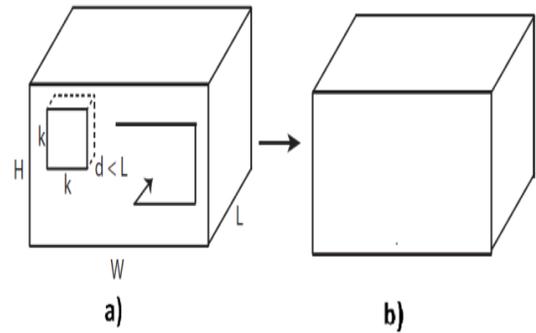


Fig. 3. Illustration of 3D Convolution Operation: a) Input Volume, and b) Output Volume.

1) *3D convolutional neural networks*: 3D-CNN is a deep learning approach [18] which is the extension of the traditional CNN with modified convolution and pooling operations. It is introduced to model the spatiotemporal features of long sequences. Sequences with long durations such as speech, videos, and EEG signals have a dependency between its segments and neglecting these dependencies may affect the robustness of recognition systems. The 3D-CNN models these temporal dependencies by applying 3D convolution operations over the input segments. In addition, the spatial correlation between pixels of video frames or different EEG channel locations can be visualized and modeled using the 3D convolution operation. The 3D-CNN has utilized for action recognition in [19].

The convolution operation is inspired by the notions of cells of the visual neuroscience [20]. The 2D-convolution operation uses 2D inputs and results in a series of 2D feature maps. Inspired by this, the 3D convolution generates a series of 3D feature volumes by processing 3D inputs, where the third dimension is the time which is modeled by consecutive input frames. From a mathematical point of view, the 3D convolution operation is calculated as follows:

$$O(x, y, z) = \sum_m \sum_n \sum_p f(m, n, p) * C(x - m, y - n, z - p) \quad (1)$$

Where O is the output of the convolution operation, f is the filter with size $m*n*p$ and C is the 3D input EEG chunk. C has usually larger size than f . The convolution is the discrete multiplication of f and C for all discrete indices x, y , and z which range from m, n , and p respectively. The 3D convolution operation is illustrated in Fig. 3: the size of the input volume is $H*W*L$. The filter size is $k*k*d$ where d is smaller than L . This results in an output volume.

2) *Network architecture*: Choosing the correct network architecture for a problem gives a better opportunity of getting more accurate results. The 3D-CNN supports a series of connected layers. Due to a large number of different layer types, it is not trivial to find an optimal chain that closely matches the given problem.

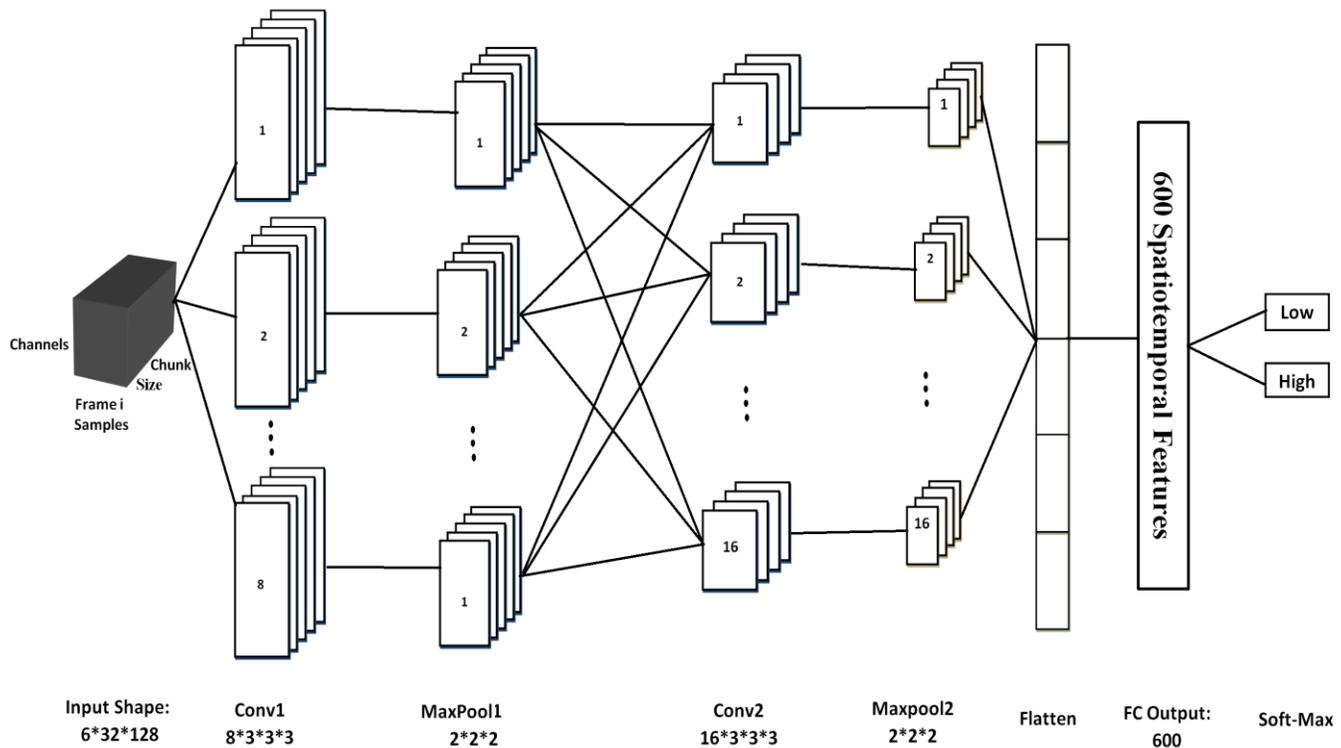


Fig. 4. Network Architecture of the Proposed 3D-CNN Model.

The adopted architecture consists of six layers. The first layer is the input volume. The middle layers are two convolution layers, each followed by a max-pooling layer. The last layer is one fully-connected layer to extract the final features. A detailed illustration of the proposed network architecture is shown in Fig. 4. For the first layer, the input volume size is 6*32*128; 6 is the number of the consecutive frames processed at once, 32 is the number of channels, and 128 is the number of samples in a frame. The kernel shape of the first convolution layer is 3*3*3: where 3, 3, and 3 are the width, height, and depth respectively. The rectified linear unit (RELU) is used as activation function in both convolution layers since it is linear, drivable, and has a simple implementation which can be expressed as:

$$\text{RELU}(x_i) = \begin{cases} x_i & x_i \geq 0 \\ 0 & x_i < 0 \end{cases} \quad (2)$$

where, x_i is the i^{th} input to the current convolution layer. The number of feature maps is set to 8. The max-pooling operation down-samples the extracted features from the convolution layer. The max-pooling layer has a resolution of 2*2*2. For the second convolution layer, the same configurations of the first convolution layer are used except for the number of feature maps which is set to 16.

Before passing the 16 resulting feature maps to the fully-connected layer, the output feature maps are reshaped to be in a vector shape. The number of output features from the fully-connected layer is 600.

IV. EXPERIMENTS AND RESULTS

Below sub-sections explains the data description, the parameter settings, the experiments, and analysis of the results.

A. Data Description

The presented system has been verified using benchmarking DEAP dataset. Using a publicly available database enables us to compare the proposed research results with the related works in literature. The DEAP dataset contains the EEG and the peripheral signals from 32 participants, and each participant watched 40 music videos each with one- minute duration. Only the EEG signals are used in the proposed work. The allowed labels in the DEAP data are valence, arousal, dominance, and liking. The subjects rated each video on a scale from 1 to 9. Only two main types of categories are tested in the proposed work: valence and arousal. Valence ranges from unpleasant to pleasant and arousal ranges from calm to active. In this paper, two binary classes for each category are tested: low and high. If the participant's rating is < 5 , the label of valence/arousal is low and if the rating is ≥ 5 , the label of valence/arousal is high.

B. Implementation Details

The proposed system works through three main steps: data augmentation, 3D input representation of the EEG signal and training and testing of the 3D-CNN model. All parameter settings are described in details in this sub-section.

1) *Training settings*: The learning rate is set to 1E-3 and the momentum is 0.9 with RMSprop optimizer. Batch size for training and testing is set to 100 samples. The K-fold cross-validation method is used to evaluate the performance of the proposed approach since it avoids using uneven dataset for testing. K is set to 5 with a true shuffle. Four folds are used for training and the remaining one fold is used for testing. The final recognition accuracy is the average over all the 5 folds. The main goals of the training process are the convergence and making the loss reaches zero. If the loss reaches zero before reaching the total number of epochs, an early stopping criterion is applied to save time processing more epochs, while the system is already converged. The proposed early stopping criterion is achieved by counting the number of times the loss reaches zero, and if this count exceeds a threshold, the optimization is stopped. This threshold is set to 3 to make sure of the system convergence. One-hundred epochs are used in the proposed experiments. For the number of features that represent the training samples of each class, the number is chosen to be 600 which is selected experimentally.

2) *Environment details*: Tensorflow framework [21] is employed in the proposed system using Core i7 device with 8Giga RAM and 960M graphics processing units (GPUs) which allowed researchers to train networks 10 or 20 times faster.

C. Pre-Processing the EEG Signals

Different pre-processing operations are applied to enhance the quality of the EEG signal and hence improve the accuracy of the emotion recognition task. The pre-processing includes performing high pass filter to get rid of any signal below 1 Hz or any dc. In addition, a band stop filter with a cutoff frequency of 50 to 60 Hz is applied to remove any unwanted noise. Besides, normalization of each channel data is performed to be between -1 and 1. The EEG signal for each video is 63 seconds. The first 3 seconds pre-trial baseline are removed from the EEG signal leaving only 60 seconds as trials for training and testing. Each the EEG signal is stationary for a small period of time [22], so, it is preferred to apply overlapping to maintain the continuity between frames. The overlap size is chosen to be 0.5.

D. Single-Label vs. Multi-Label Emotion Recognition

The most well-known approach to classify an input instance into valence and arousal class labels is to simply train an independent classifier for each label at once. This is called single-label classification (SLC). Multi-label classification (MLC) [23] aims to classify instances where each instance belongs to more than one class simultaneously. In emotion recognition case, MLC intends to classify input instance into its four combinations of valence and arousal; low valence-low arousal, low valence-high arousal, high valence-low arousal,

and high valence-high arousal. MLC saves time processing each dimensional label in separate.

In the proposed work, a binary representation is associated for each input instance to represent its label. In the case of SLC, only two digits are required to represent the two classes of valence/arousal (low and high) such that 10 mean high valence/arousal and 01 mean low valence/arousal. For MLC case, the label of each input instance is represented in four digits to express its four combinations. The first two digits represent the labels of valence and the last two digits represent the labels of arousal. For example, an input instance with label 1001 means that input instance is classified as low valence and high arousal simultaneously. In the proposed work, single-label and multi-label experiments are conducted to investigate the effectiveness of each methodology on the emotion recognition performance.

E. Results and Discussions

To show the effectiveness of the proposed system, a set of experiments are conducted using the DEAP data. Each experiment is implemented using the best configuration achieved till now from its previous experiment.

1) *Single-label EEG based emotion recognition*: In this experiment, valence labels are classified in separate from arousal labels. A flag with two values “valence” or “arousal” is set. If the flag is “valence”, the input sample is classified as low/high valence. If the flag is “arousal”, the input sample is classified as low/high arousal. Two experiments are conducted: the first is to choose the appropriate chunk size and the second is to increase the number of training samples.

a) *Choosing the appropriate chunk size*: chunk size is the number of consecutive frames that are combined together in one chunk as an input to the 3D-CNN method to model the temporal dependency between the EEG signal segments. The duration of each frame is one second, so the chunk size refers to the required number of seconds to describe the given emotion. Table I shows the experimented chunk sizes and the corresponding average accuracy over all users.

As illustrated, the accuracy increases as the chunk size increase until a specific range and reaches its maximum using 6 seconds. This concludes that the 3D-CNN needs about 6 seconds to give a precise decision about the input emotion. As long as the data from each video has 60 seconds trial and the 60 seconds are divided into 1s frames, 60 frames are achieved. Appending every 6 successive frames together as a chunk, every video contains 10 chunks. Since the overlap size is 0.5, 20 chunks are achieved for each video on average. Hence, the total readings are 25376 chunks for the 32 users and 40 videos.

TABLE I. AVERAGE ACCURACY FOR VALENCE AND AROUSAL USING DIFFERENT CHUNK SIZES

	Chunk Size				
	3	5	6	7	9
Valence	70.80	71.97	76.3	74.77	73.98
Arousal	72.60	74.62	78.10	75.80	75.22

b) *The effect of augmentation phase:* One of the limitations of machine learning methods is the availability of sufficient training data to get high recognition performance [24]. The EEG data has only 40 one-minute videos for each subject. To solve this limitation, several noisy versions of the same video signal are generated to increase the number of video samples. The augmentation process is applied to the training samples only and clean data is used for testing. The type of the noise added to the original signal is the white Gaussian noise with zero mean and unit variance. In order not to affect the quality of original signal and hence the accuracy of the system, the signal-to-noise ratio (SNR) between original the EEG signal and the noise signal is set to a small value which is 5.

Fig. 5 illustrates a comparison between the accuracy values with and without augmentation. This experiment is done using a chunk size of 6 frames with 0.5 overlap and pre-processing operations as mentioned in section C. Several numbers of augmentation files have experimented; 10, 30, and 50. The first element in the x-axis in Fig. 5 refers to no augmentation case. As clear, the average accuracy increases with the increase in the number of augmentation files since having bigger data allows the system to generalize. And the results indicate that better results can be achieved by increasing the number of noisy files per video.

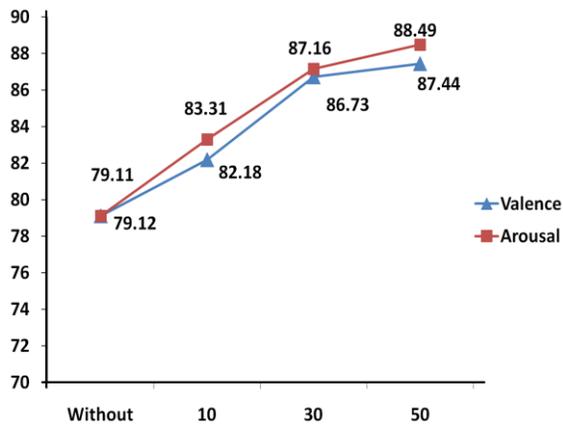


Fig. 5. The Effect of Augmentation on Average Accuracy Overall.

c) *Comparative Study:* the proposed work is compared to the state of the art EEG-based emotion recognition methods that worked on DEAP data. The comparison is presented in Table II.

TABLE II. COMPARISON WITH THE RELATED WORKS IN LITERATURE AND THE PROPOSED METHOD IN TERMS OF ACCURACY

	Valence	Arousal
Yoon and Chung [7]	70.9	70.1
Naser and Saha [8]	64.3	66.2
Atkinson and Campos [9]	73.41	73.06
Chen et al. [11]	73	75.63
Koelstra et al. [12]	57.6	62
Rozgic et al. [13]	76.9	68.4
Li et al. [10]	72.06	74.12
Alhagry et al. [14]	85	85
The proposed method	87.44	88.49

Yoon and Chung [7] extracted FFT from the EEG segments, Naser and Saha [8] used DT-CWPT for feature extraction. DT-CWPT has less accuracy than FFT which indicates the superior effectiveness of FFT in emotion recognition. Atkinson and Campos [9] computed a set of statistical features besides some frequency bands features. The proposed method in [9] is better than both works in [7], [8] due to the use of a different set of features. Their work still less than the proposed method since their features are hand-crafted. Hand-crafted features require a huge amount of engineering skill and domain expertise to select the best set of features that best represent input data.

Chen et al. [11] computed the power values of 6 frequency bands for each electrode as features and got a recognition accuracy of 73 and 75.63 for valence and arousal classes respectively. Even though they are using hand-crafted features, their accuracy is a bit high due to working on only 10 participants. Both Rozgic et al. [13] and Koelstra et al. [12] extracted the same features from the EEG signals. However, their accuracies still less than the proposed method since they are using hand-crafted features (power of frequency bands). This proves the superiority of deep learning features especially the 3D-CNN features for EEG-based emotion recognition task.

Li et al. [10] extracted spatiotemporal features from two different architectures (CNN with RNN) combined stacked together. While CNN and RNN got good recognition accuracy, they are two different models. The CNN used stacked layers of convolution operations, and the RNN used gated cells called Long Short-term Memory (LSTM). On the other hand, the 3D-CNN extracts the spatiotemporal features in one- end-to-end architecture with sharing parameters. Alhagry et al. [14] have a larger accuracy compared to other works in literature since they used the RNN which is a deep learning method the models the time variations in the EEG signal. Although RNN has promising results in emotion recognition tasks, RNN has the limitation of increasing depth while extraction with the spectral and temporal features caused by the great number of parameters. And, it extracts only temporal features and it requires another method such as CNN to model the spatial variations in the EEG signal. However, the 3D-CNN is a compact model that is able to model the spatiotemporal variations simultaneously.

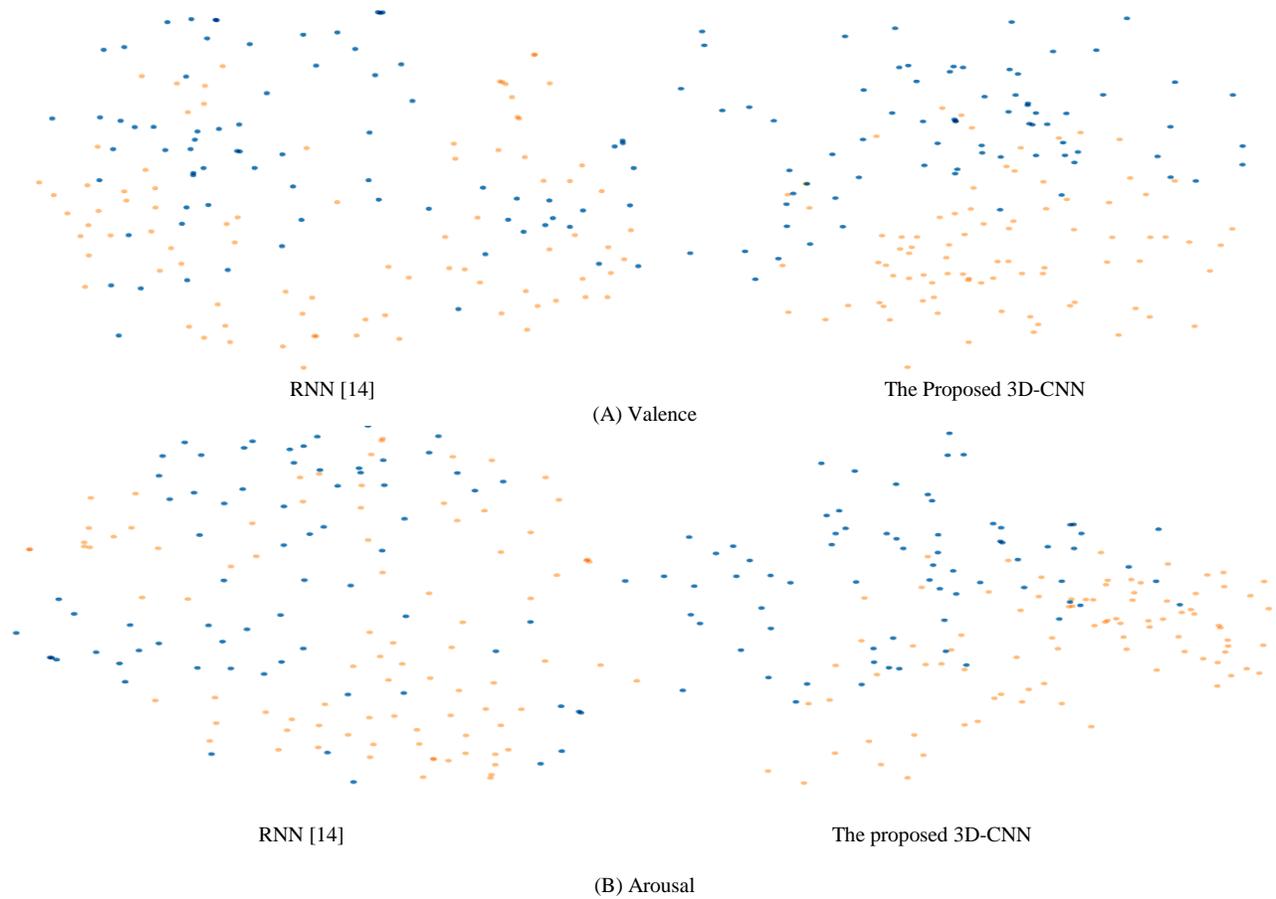


Fig. 6. A t-SNE Visualization Of Testing Samples for User 1 for both Valence (A) and Arousal (B) from Two Different Models: the RNN [14] (left) and the Proposed the 3D-CNN (right). Blue Samples Belong to the Low Class and High-Class Samples are Marked with Yellow. Best Viewed in Color.

The 3D-CNN significantly outperforms recent-related methods in the literature. One of the main reasons for the superiority of the 3D-CNN is the existence of the 3D-convolution operation in its architecture. One of the main advantages of using the convolution operation is the parameter sharing since one feature detector of one part of input can be useful in another part of the same input. In addition, the convolution operation has the advantage of the sparsity of connections since each output value depends on a small number of inputs. These advantages provide the 3D-CNN with the capability of extracting representative spatial features that visualize the correlation between different channel locations. The 3D-convolution operation works over consecutive frames which help in extracting temporal features from the consecutive EEG segments.

Table III describes a set of different performance measures for the proposed system. All the values indicate the effectiveness of the 3D-CNN for emotion recognition using brain signals.

TABLE III. DIFFERENT PERFORMANCE MEASURES FOR THE PROPOSED METHOD

	Sensitivity	Specificity	Precision	F1-Score
Valence	0.88	0.88	0.83	0.86
Arousal	0.85	0.91	0.88	0.86

To figure out the effectiveness of the 3D-CNN features in discriminating between dimensional labels, the t-SNE tool is used to visualize the test samples of one user as shown in Fig. 6 in comparison with best recent work [11]. As illustrated in Fig. 6, the proposed the 3D-CNN features results in high discriminated classes due to its compact model which do spatial and temporal classification in one end-to-end model. These results show that the proposed the 3D-CNN method produces discriminative feature representations which could give accurate emotion recognition system.

2) *Multi-label EEG-based emotion recognition*: In emotion recognition, classifying the input emotion in terms of only one dimension at a time leads to an incomplete description of given input emotional video. For example, for a sample to be happy in the dimensional emotions as in Fig. 7; it must be in high arousal and high valence. Hence, classifying an input based on valence only or arousal only would not give a complete description of an emotion.

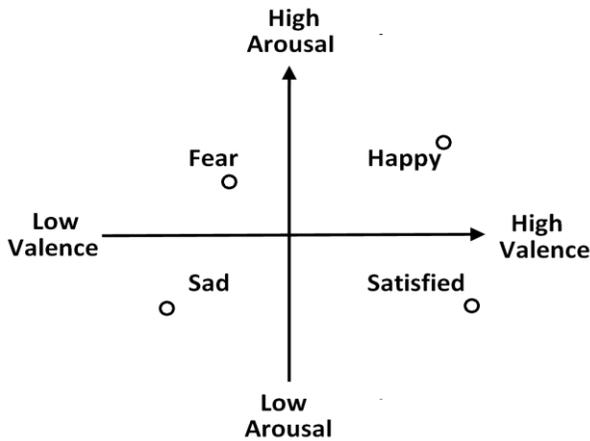


Fig. 7. Illustration of Dimensional Emotions: Valence and Arousal.

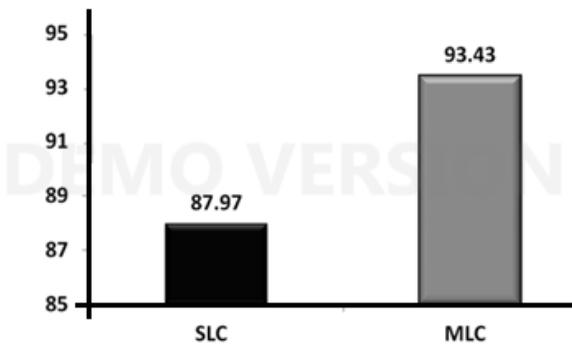


Fig. 8. Comparison of the Proposed SLC and MLC Methods.

Sigmoid activation function which is commonly used in multi-label classification problems [25], assumes no dependency between class labels. Softmax activation function is chosen in the proposed work of multi-label classification since there is a dependency between dimensional labels. For example, one input instance could not be classified as 1100; which is low and high valence simultaneously and not any of the arousal labels. Softmax activation function can be expressed as:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (3)$$

Where z_i is the i^{th} weighted input instance after passing through the last layer in the 3D-CNN model, and K is the total number of samples in the last layer.

In the proposed method, the argmax function is applied to get the label at which the maximum probability occurs. It is first applied twice, one time for the first two digits of binary representation to get the label of maximum prediction probability of valence labels, and another time for the second two digits to get the maximum prediction probability of arousal labels. Then, the mean of correct predictions is calculated for both valence and arousal classes.

The average accuracy per valence and arousal can be calculated simply by adding the two means of valence and arousal, then, the result is divided by 2.

Fig. 8 compares the proposed SLC and MLC methods where SLC is the best result achieved in experiment 2. The comparison indicates the ability of the 3D-CNN to model the correlation between valence and arousal labels. In addition considering the correlation between valence and arousal gives a higher performance which can reach 93.43.

V. CONCLUSION

In this paper, the 3D-CNN emotion recognition approach is proposed to extract the spatiotemporal features to model the temporal dependencies between the EEG signals. Since the 3D-CNN requires 3D inputs, a novel method that represents the EEG signals into a 3D format from multi-channel signals has been developed. The frame samples from multi-channels are used to create a 2D spatial matrix. Then, the time dimension is appended by concatenating m consecutive frames together to form the 3D input volume. In order to show the effectiveness of the proposed method, the DEAP data is used. Since most of the publicly available EEG datasets have fewer amounts of data per subject, the data augmentation phase is employed to increase the number of samples per subject by adding noise signals to the original EEG signals.

It has been shown from the experimental work that the proposed method is capable of producing a very high recognition accuracy compared to works in literature in the same domain. From this comparative study, the proposed approach is capable of achieving a significant improvement in emotion recognition from the EEG signals. In addition, the 3D-CNN proves its superiority in visualizing the correlation between valence and arousal labels and hence gives promising recognition accuracy. The advantage of using the 3D-CNN is the ability to extract spatial and temporal features in one end-to-end model. In addition, it works well using time domain raw signals to construct frames for feature learning. Also, it does not need a very deep architecture to work with and hence a less processing time. Future work will include combining different modalities together with the EEG signals such as the face or the eye to increase the performance of the spatiotemporal feature learning in emotion recognition systems.

REFERENCES

- [1] N.N. Khatri, Z.H. Shah, and S.A. Patel, "Facial expression recognition: A survey," International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp. 149–152, 2014.
- [2] Q. Yao, "Multi-sensory Emotion recognition with speech and facial expression," Ph.D. dissertation, Computing, University of Southern Mississippi, 2014.
- [3] P. Ekman, and W. V. Friesen, Unmasking the face: a guide to recognizing emotions from facial expressions, 1st ed., Englewood Cliffs, N. J: Prentice Hall, 1975.
- [4] C. Brunner, C. Vidaurre, M. Billinger, and C. Neuper, "A comparison of univariate, vector, bilinear autoregressive, and band power features for brain-computer interfaces," Medical and Biological Engineering and Computing, vol. 49, no. 11, pp. 1337–1346, 2011.
- [5] J. Kim, and E. Andre, "Emotion recognition based on physiological changes in music listening," in Proceedings of IEEE International Conference on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, 2008, pp 2067–2083.
- [6] G. K. Verma, and U. S. Tiwary, "Multimodal fusion framework: a multi-resolution approach for emotion classification and recognition from physiological signal," NeuroImage, vol. 102, pp. 162–172, 2014.

- [7] H. J. Yoon and S. Y. Chung. "EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm," *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2230–2237. 2013.
- [8] D. S. Naser and G. Saha. "Recognition of emotions induced by music videos using DT-CWPT," 2013 Indian Conference on Medical Informatics and Telemedicine, India, 2013, pp. 53–57.
- [9] J. Atkinson and D. Campos. "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Systems with Applications*, vol. 47, pp. 35–6, 2015.
- [10] K. Li, et al., "Emotion recognition from multi-channel the EEG data through convolutional recurrent neural network," in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*, Shenzhen, China, 2017, pp. 352–359.
- [11] J. Chen, B. Hu, L. Xu, P. Moore, and Y. Su, "Feature-level fusion of multimodal physiological signals for emotion recognition," in *Proceedings of IEEE International Conference Bioinformatics and Biomedicine*, Washington, USA, 2015, pp. 395–4.
- [12] S. Koelstra, C. Muhl, and M. Soleymani, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–13, 2012.
- [13] V. Rozgic, S. N. Vitaladevuni, and R. Prasad, "Robust the EEG emotion classification using segment level decision fusion," In *2013 IEEE Conference of Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013, pp. 1286–1290.
- [14] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on the EEG using LSTM recurrent neural network," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 355-358, 2017.
- [15] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal recurrent neural network for emotion recognition," *Journal of Latex Class Files*, vol. 13, no. 9, pp. 1–8, 2017.
- [16] K. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Neural Information Processing Systems*, pp. 1106–1114, 2012.
- [17] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of Computer Vision and Pattern Recognition*, 2014, pp. 1409-1556.
- [18] D. Maturana, and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, Hamburg, Germany, 2015.
- [19] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," *The IEEE International Conference on Computer Vision*, 2015, pp. 4489-4497.
- [20] D. H. Hubel, and T. N. Wiesel, "Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex," *Journal of Physiology*, vol. 160, pp. 106–154, 1962.
- [21] M. Abadi, P. Barham, J. Chen, and Z. Chen, A. Davis, J. Dean, et al, "TensorFlow: A system for large-scale machine learning," in the *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, GA, USA, 2016, pp 265–283.
- [22] N. Hazarika, J. Z. Chen, A. C. Tsoi, and A. Sergejew, "Classification of the EEG signals using the wavelet transform." in *Proceedings of the 13th International Conference in Digital Signal Processing*, Santorini, Greece, 1997, pp. 61-72.
- [23] G. Tsoumakas, I. Katakis, and I. Vlahavas, "A review of multi-label classification methods," in *Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery*, 2006, pp. 99–109.
- [24] Y. Zhang, X. Ji, and S. Zhang, "An approach to EEG-based emotion recognition using combined feature extraction method," *Neuroscience Letters*, vol. 633, pp. 152–157, 2016.
- [25] L. Lenc, and P. Král, "Combination of neural networks for multi-label document classification," *International Conference on Applications of Natural Language to Information Systems*, 2017, vol 10260, pp. 278-282.

Enhanced and Improved Hybrid Model to Prediction of User Awareness in Agriculture Sector

A.V.S. Pavan Kumar

Research Scholar,
Department of Computer Science and Engineering,
GIT, GITAM, Visakhapatnam

Dr. R. Bhramaramba

Associate Professor,
Department of Information Technology,
GIT, GITAM, Visakhapatnam

Abstract—Agriculture is the backbone of Indian economy and is the main income source for most of the population in India. So farmers are always curious about yield prediction. Crop yield depends on various factors like soil, weather, rain, fertilizers and pesticides. Several factors have different impacts on agriculture, which can be quantified using appropriate statistical methodologies. Applying such methodologies and techniques on historical yield of crops, it is possible to obtain information or knowledge which can be helpful to farmers and government organizations for making better decision and policies which lead to increased production. The main drawbacks of Indian farmers are they do not have proper knowledge regarding crop yield based on soil necessities. So in this paper, we proposed and developed an Improved Hybrid Model (which is combination of both classification, i.e. Artificial Neural Networks and clustering approach i.e. k-means (works based on Euclidean distance)) to provide awareness, usage and prediction to each farmer that relates to classify different crop yield representation based on soil necessity. For that we collected farmer's data from standard repositories like http://www.tropmet.res.in/static_page.php?page_id=52#data and then using that data provide awareness and other parameter sequences to all the farmers in India. Our experimental results show efficient e-agriculture with respect to user awareness, usage and prediction with respect to prediction, recall and f-measure for supporting real time marketing of different agriculture products.

Keywords—Agriculture products; e-agriculture; classification; clustering; ensemble model

I. INTRODUCTION

India is described by little homesteads. More than 75% of aggregate land capitals inside the nation are under 5 sections of land. Most yields are rain sustained, with pretty much 45% of the land inundated. According to a few estimations, around 55% of aggregate populace of India relies upon cultivation. In the US, in light of the fact of overwhelming automation of agriculture, it is around 5%. India is one of the greatest makers of agrarian items and still has exceptionally less ranch profitability. Efficiency should be expanded so agriculturists can get more pay from a similar land parcel with less work. Accuracy in agriculture gives an approach to do it. Exactness cultivating, as the name suggests, alludes to the applying of exact and appropriate aggregate of remark like pee, manures, soil and so on at the best possible time to the gizzard for expanding its profitability and expanding its yields. Not all exactness agriculture frameworks offer best outcomes [7], [9]. In any case, in agribusiness it is vital that the proposals made

are exact and exact in light of the fact that if there should be an occurrence of mistakes it might prompt overwhelming material and capital misfortune [13]. Numerous inquiries about are being done, so as to achieve an exact and proficient model for trim forecast.

Improved Hybrid Model (which is combination of both classification i.e., Artificial Neural Networks and clustering approach i.e. k-means (works based on Euclidean distance)) (Ensemble approach) is one such strategy that is incorporated into such research works. Among these different machine learning procedures that are being utilized as a part of this field. This paper proposes a framework that uses the voting technique to assemble a productive and exact model. The agriculture part in India is at present confronting a troublesome stage. India is moving towards an agribusiness crisis because of deficient interest in irrigational and farming framework, absence of consideration, insufficient land administration, not given of reasonable costs to ranchers for their yields and inadequate land change in India, and so on. Sustenance creation and efficiency in India is declining while its nourishment utilization is expanding. The circumstance has additionally been exacerbating because of utilization of sustenance grains as a result of interest of bioenergizers. As India does not have ports and calculated frameworks for extensive - scale sustenance imports, the arrangement of import of sustenance grains would be difficult. In our proposed approach e-agriculture is a rising field in the association of agricultural informatics, advancement and enterprise which is focusing to farming administrations, innovation dissemination and data conveyed or created through the Internet and related advances [8]. In particular, it connects with the conceptualization, outline, improvement, appraisal and application of imaginative approaches to utilize dynamic or developing Information and Communication Technologies (ICTs) [12].

Our proposed approach is a rising technology for upgrading existing agriculture [10], [11] and sustenance security through improved procedures for learning access and to switch to utilizing data and correspondence advancements. The World Summit on the Information Society (WSIS) Plan of Action includes e-Agriculture as a locale of capacity of data and correspondence innovations (ICTs). In short e-Agriculture will associate every concerned individual beginning from ranchers to scientists together. Agriculturists can get the coveted data at any moment of time from any piece of world and they can

likewise get the assistance from specialists seeing their concern instantly by without moving anyplace.

II. REVIEW OF RELATED WORK

Satish et al. [1] states the necessities and arranging required for building up a product display for accurate cultivation is examined. It profoundly ponders the nuts and bolts of accurate cultivation. The creator's beginning from the nuts and bolts of accurate cultivation and moves towards building up a model that would bolster it. This paper depicts a model that applies Precision Agriculture (PA) standards to little, open ranches at the individual rancher and yield level, to influence a level of control over changeability. The extensive target of the model is to convey guide warning administrations to even the littlest rancher at the level of his/her littlest plot of harvest, utilizing the most open innovations, for example, SMS and email. This model has been intended for the situation in Kerala State where the normal holding size is much lower than the vast majority of India. Thus this model can be situated somewhere else in India just with a few alterations. Anshal Savla et al. [2] makes a qualified meditation of grouping calculations and their execution in yield forecast in exactness farming. These calculations are executed in an informational collection gathered for quite a while in yield expectation on soya bean trim. The calculations utilized for yield forecast in this paper are Support Vector Machine, Random Forest, Neural Network, REPTree, Bagging, and Bayes. The conclusion drawn at the end is that packing is the best calculation for yield expectation among the above expressed calculations since the mistake deviation in stowing is least with a mean outright blunder of 18985.

M.P. Singh et al. [3] demonstrates the significance of yield choice and the elements choosing the harvest choice like creation rate, showcase cost and government strategies are talked about. This paper proposes a Crop Selection Method (CSM) which takes care of the product determination issue and enhances net yield rate of the harvest. It recommends a progression of yield to be chosen over a season considering factors like climate, soil write, water thickness, edit type. The anticipated estimation of powerful parameters decides the precision of CSM. Consequently there is a need to incorporate a forecast strategy with enhanced exactness and execution. Liying Yang et al. [4] expects to tackle the pivotal issue of choosing the classifiers for the troupe learning. A technique to choose a best classifier set from a pool of classifiers has been proposed. The proposition expects to accomplish higher precision and execution. A technique called SAD was proposed in view of precision and characterization execution. Utilizing Q measurements, the reliance between most significant and exact classifiers is distinguished. The classifiers which were not picked were joined to shape the troupe. This measure should guarantee higher execution and decent variety of the outfit. Different strategies, for example, SA (Selection by Accuracy), SAD (Selection by precision and Diversity) and NS (No choice) calculation were distinguished. At last it is gathered that SAD works superior to others.

Shakil Ahamed et al. [5] propose different order techniques to group the liver ailment informational index. The paper stresses the requirement for exactness since it relies upon the

dataset what's more, the realizing calculation. Order calculations such as Naïve Bayes, ANN, ZeroR and VFI were utilized to group these sicknesses and think about the adequacy, adjustment rate among them. The execution of the models was contrasted with precision and computational time. It was reasoned that every one of the classifiers with the exception of innocent bayes demonstrated enhanced prescient execution. Multi-layer perception demonstrates the most noteworthy precision among the proposed calculations. Aymen E Khedr et al. [6] tries to take care of the issue of nourishment frailty in Egypt. It proposes a system which would foresee the generation, and import for that specific year. It utilizes Artificial Neural Networks alongside Multi-layer perceptron in WEKA to fabricate the forecast. Toward the finish of the procedure we would have the capacity to imagine the measure of generation import, need and accessibility. Thus it would settle on choices on whether sustenance must be additionally transported in or not. The dirt datasets in paper are examined and a classification is anticipated. From the anticipated soil class the edit yield is distinguished as a Classification run the show. Credulous Bayes furthermore, KNN calculations are utilized for trim yield expectation. The future work expressed is to make productive models utilizing different characterization methods, for example, bolster vector machine, main segment investigation. The advantages of the Indian agriculturists: if the market and climate data is conveyed to their cell phones [14]. Furthermore, this has been led with a randomized trial in 100 towns of Maharashtra. This administration has been sent in by a business benefit called Reuters Market Light (RML). The treated ranchers connect RML data with various choices they have made in the farming, and we find that the treatment influenced spatial arbitrage and product reviewing. Be that as it may, the size of these impacts is little. We discover no measurable noteworthy normal impact of treatment on the cost got by ranchers, edit esteem - included, trim misfortunes coming about because of rainstorms, or the probability of changing harvest assortments and development hones [15]. The information that are applicable of the required quality dependably have the capability of expanding effectiveness in all circles of movement of an Indian rancher, accordingly the developing situation of the deregulated horticulture, has brought a need and direness to guarantee it in a fundamental piece of basic leadership. In this way, investigating IT as a key device is the advantage of country like India of accepted significance. Here the data meets the Indian ranchers when all is said is done which are recorded broadly.

III. DESCRIPTION OF PROBLEM

The farming industry in Native Indian is currently experiencing a hard stage. Native Indian is moving towards an farming emergency due to inadequate investment in irrigational and farming facilities, lack of attention, worthless area management, non-given of fair prices to farm owners for their plants and inadequate area change in Native Indian, etc. Meals manufacturing and efficiency in Native Indian is decreasing while its food consumption is increasing. The situation has further been difficult due to use of food grain because of demand of bioenergy sources. As Naive Bayesian does not have slots and logistical systems for large - scale food

imports, the solution of transfer of food grain would be a challenge.

By the use of ICT, India's food manufacturing and efficiency has been increased for farming reasons. The developed countries are using technological innovation of laser in place of vehicles to plough lands. This helps in improving the use of a range of information parameter such as water, plant seeds, plant foods, etc. The issue occurs here is that Native Indian farm owners cannot pay for this technological innovation. In addition, energy and power also cause a major issue for Native Indian farm owners and choice of energy like solar panel technology sections, controlled and enhanced by ICT.

IV. PROPOSED METHODOLOGY AND IMPLEMENTATION

The proposed technique, i.e. Enhanced Hybrid Model performs information parceling with Principal segment. It parcels the given informational index into k sets. The middle of each set can be utilized as great beginning group focuses and afterward allot every datum focuses to its closest bunch centroid. The underlying centroids of the bunches are given as information. It begins by framing the underlying bunches in view of the relative separation of every datum point from the underlying centroids. The Euclidean separation is utilized for deciding the closeness of every datum point to the group centroids. For every datum indicate, the bunch which it is doled out and its separation from the centroid of the closest group are noted. For each group, the centroids are recalculated by taking the mean of the estimations of its information focuses. The strategy is relatively like the first k-implies calculation aside from that the underlying centroids are figured deliberately. The following stage is an iterative procedure which makes utilization of a heuristic technique to enhance the proficiency. Amid the emphasis, the information focuses may get redistributed to various bunches. The strategy includes monitoring the separation between every datum point and the centroid of its present closest group. Toward the start of the emphasis, the separation of every datum point from the new centroid of its present closest bunch is resolved. On the off chance that this separation is not exactly or equivalent to the past closest separation, that means that the information point remains in that group itself and there is no compelling reason to process its separation from different centroids. This outcome in the sparing of time required to register the separations to k-1 group centroids. Then again, if the new centroid of the present closest group is more inaccessible from the information point than its past centroid, there is a shot for the information point getting incorporated into another closer bunch. All things considered, it is required to decide the separation of the information point from all the group centroids. This technique enhances the proficiency by lessening the quantity of calculations.

This ANN K-Means clustering approach illustrated as follows: Several steps of data processing were followed by details pre-processing, data reduction, Data Mining, details clustering, information interpretation, neighborhood study and statistical diagnosis shown in Fig. 1.

Algorithm 1: Procedure for proposed algorithms with respect to different attributes

Algorithm	: ANN and K-Mean clustering algorithm
INPUT	: Agriculture Data set with d dimensions
OUTPUT	: K number of Clusters
Steps	
1.	Reduce the D dimension of the N data using Artificial Neural Networks (ANN) and prepare another N data with d dimensions ($d < D$).
2.	The Principal components are ordered by the amount of variance.
3.	Choose the first principal component as the principal axis for partitioning and sort it in ascending order.
4.	Divide the Set into k subsets where k is the number of clusters.
5.	Find the median of each subset.
6.	Use the corresponding data points for each median to initialize the cluster centers
7.	Compute the distance of each data-point x_i ($1 \leq i \leq n$) to all the centroids c_j ($1 \leq j \leq k$) using Euclidean distance formula.
8.	For each data object x_i , find the closest centroid c_j and assign x_i to the cluster with nearest centroid c_j and store them in array Cluster[] and the Dist[] separately. <ul style="list-style-type: none"> a. Set Cluster[i] = j, j is the label of nearest cluster. b. Set Dist[i] = $d(x_i, c_j)$, $d(x_i, c_j)$ is the nearest Euclidean distance to the closest center.
9.	For each cluster j ($1 \leq j \leq k$), recalculate the centroids;
10.	Repeat
11.	for each data-point <ul style="list-style-type: none"> 11.1 Compute its distance from the centroid of the present nearest cluster 11.2 If this distance is less than or equal to the previous nearest distance, the data-point stays in the cluster
12.	Else
13.	For every centroid c_j
14.	Compute the distance of each data object to all the centre
15.	Assign the data-point x_i to the cluster with nearest centroid c_j
16.	For each cluster j ($1 \leq j \leq k$), recalculate the centroids;
17.	Until the convergence criteria is met.

Performance evaluation for different crop yield parameters are like pH, TA, TDS, MG and CA with feasible environments sequences with respect to soil parameters shown in Table I. Following table shows effective results of classification measurements like precision, f-measure and recall with data representation may appear effective utilization soil analysis in agriculture sector for real time applications.

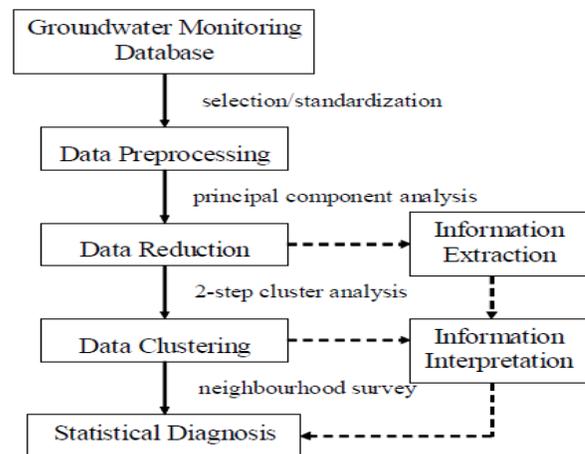


Fig. 1. Step by Step Procedure of Proposed Approach.

TABLE I. BASIC PARAMETER SEQUENCES FOR DIFFERENT SOIL PARAMETERS

Performance Measures	Accuracy	Precision	Recall	F-Measure
PH	0.95	0.83	0.84	0.84
TA	0.89	0.88	0.82	0.81
TDS	0.87	0.86	0.81	0.85
MG	0.85	0.89	0.84	0.84
F	0.87	0.76	0.84	0.82
CA	84.59	0.85	0.85	0.78

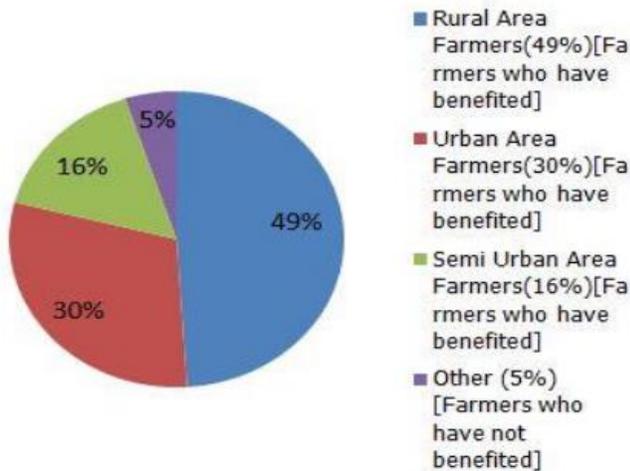


Fig. 2. Information of Different Soil Parameters at Different Areas in India.

Based on these parameters, performance evaluation of different formers at different parameter sequences is shown in Fig. 2.

A. Data Collection

The data set containing the ground specific features which are gathered from Poly test Labs ground examining lab, Pune, Maharashtra, Indian. In addition, similar resources of general vegetation information were also used from Marathwada University. The vegetation regarded in our design includes groundnut, impulses, pure cotton, vegetables, bananas, paddy, sorghum, sugarcane, and cilantro. The number of illustrations of each vegetation available in the training information set is proven. The attributes regarded where Detail, Structure, Ph, Soil Shade, Permeability, Water flow and drainage, Standard water having and Break down. The above mentioned factors of ground play a major part in the crop's capability to eliminate water and nutritional value from the ground. For vegetation development to be possible, the ground must provide appropriate atmosphere for it. Soil is the core of the origins. The water having potential decides the crop's capability to process nutritional value and other nutritional value that are turned into ions, which is the form that to obtain can use. Structure decides how permeable the ground is and the convenience of air and water action which is essential to prevent the vegetation from becoming water

logged. The stage of acid or alkalinity (Ph) is expert varying which affects the accessibility of ground nutritional value. The action of microorganisms present in the ground and also the stage of exchangeable metal can be impacted by PH. The water holding and drainage figure out the infiltration of origins. Hence for the following reasons the above mentioned parameters are viewed for selecting vegetation.

V. EXPERIMENTAL RESULTS

The suggested program can be experimentally confirmed in conditions of clustering efficiency. Evaluation can be made centered on parameters such as Time and Precision. A past method of E-agriculture does not use any information exploration methods. The proposed program uses information exploration way of clustering strategy to team the data of farm owners. The clustering efficiency can be calculated in following conditions namely precision, recall, F-measure.

Precision value which is calculated relies on the recovery of information at real beneficial forecast, incorrect beneficial. In health care information precision is calculated as the portion of good outcomes came back that are appropriate and shown in Fig. 3.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

TP-True positive FP-true negative

Recall value is measured relies on the recovery of details at real beneficial forecast, incorrect adverse shown in Fig. 4. In healthcare details precision is measured the amount of beneficial outcomes came back that are Remember in this perspective is also known to as the True Positive Amount. Remember is the portion of appropriate circumstances that are recovered,

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

FN – false negative

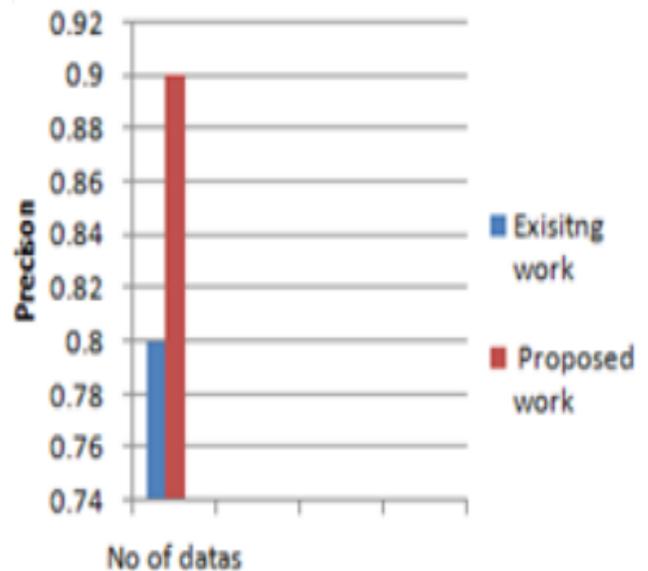


Fig. 3. Precision Comparison for Different Soil Ratios.

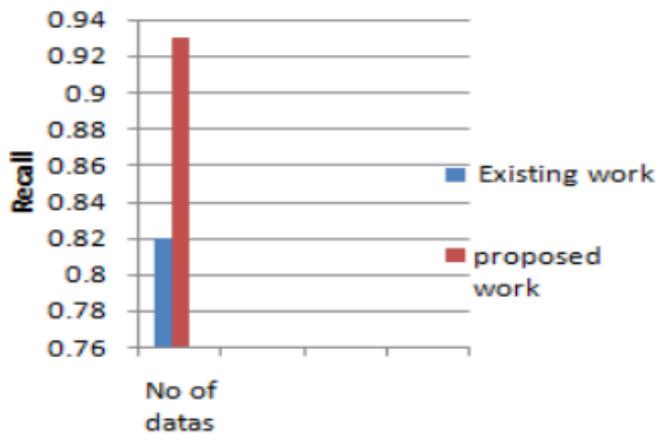


Fig. 4. Recall Comparison for Different Soil Parameters.

The F- Measure computes some average of the information retrieval precision and recall metrics is shown in Fig. 5.

Total comparison results for different soil parameters present in agriculture crop yield areas for real time applications is shown in Table II and Fig. 6.

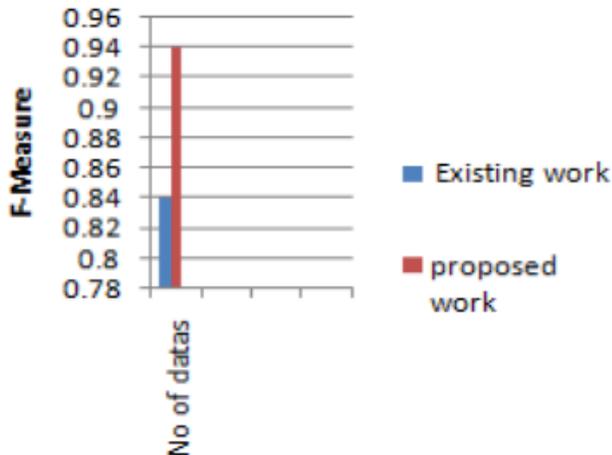


Fig. 5. F-Measure Parameter Sequences for Different Soil Parameters.

TABLE II. DIFFERENT VALUES FOR PROPOSED MEASUREMENTS FOR DIFFERENT SOIL QUALITY PARAMETERS

Soil Quality Parameters						
Performance Measures	PH	TA	TDS	MG	F	CA
Accuracy	80.25	83.42	83.89	82.38	81.29	84.39
Precision	0.83	0.88	0.85	0.83	0.70	0.82
Recall	0.84	0.82	0.81	0.84	0.84	0.85
F Measure	0.84	0.81	0.85	0.84	0.82	0.78

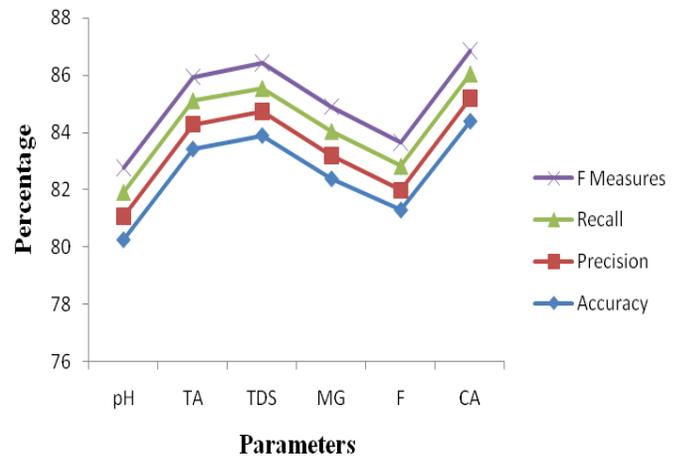


Fig. 6. Proposed Approach Performance at Different Soil and Water Quality Parameters.

Above results shown in Table II was proposed approach performance with different soil and other representations.

VI. CONCLUSION

Majority of farm owners in the state or nation are not aware that real time servers can be used to outperform companies and get details. The government should also perform sensitization to make attention for the farm owners on how best they can use information technological innovation to perform agribusiness. Our perform would help farm owners to increase efficiency in farming, avoid ground deterioration in harvested area, and decrease substance use in plants manufacturing and effective use of water sources. Our upcoming work is targeted at an enhanced information set with large number of features and also utilizes how to generate prediction.

REFERENCES

- [1] Babu, Satish. "A software model for precision agriculture for small and marginal farmers." In Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS), 2013 IEEE, pp. 352-355. IEEE, 2013.
- [2] Savla, Anshal, Nivedita Israni, Parul Dhawan, Alisha Mandholia, Himtanaya Bhadada, and Sanya Bhardwaj. "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture." In Innovations in Information, Embedded and Communication Systems (ICIECS), 2015 International Conference on, pp. 1-7. IEEE, 2015.
- [3] Kumar, Rakesh, M. P. Singh, Prabhat Kumar, and J. P. Singh. "Crop Selection Method to maximize crop yield rate using machine learning technique." In Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on, pp. 138-145. IEEE, 2015.
- [4] Yang, Liying. "Classifiers selection for ensemble learning based on accuracy and diversity." Procedia Engineering 15 (2011): 4266-4270.
- [5] Ahamed, AT M. Shakil, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman, and Rashedur M. Rahman. "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh." In 2015 16th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 1-6. IEEE, 2015.

- [6] Khedr, Ayman E., Mona Kadry, and Ghada Walid. "Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector applied case on food security information center ministry of agriculture, Egypt." *Procedia Computer Science* 65 (2015): 633-642.
- [7] Paul, Monali, Santosh K. Vishwakarma, and Ashok Verma. "Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach." In *Computational Intelligence and Communication Networks (CICN)*, 2015 International Conference on, pp. 766-771. IEEE, 2015.
- [8] Sharma, Latika, and Nitu Mehta. "Data mining techniques: A tool for knowledge management system in agriculture." *international journal of scientific & technology research* 1, no. 5 (2012): 67-73.
- [9] Mucherino, Antonio, and Georg Ruß. "Recent Developments in Data Mining and Agriculture." In *Industrial Conference on Data Mining-Workshops*, pp. 90-98. 2011.
- [10] Ruß, G., 2009, July. "Data mining of agricultural yield data: A comparison of regression models". In *Industrial Conference on Data Mining* (pp. 24-37). Springer, Berlin, Heidelberg.
- [11] Rajesh, D. "Application of spatial data mining for agriculture." *International Journal of Computer Applications* 15, no. 2 (2011): 7-9.
- [12] Darcy Miller, Jaki McCarthy, Audra Zakzeski, "A Fresh Approach to Agricultural Statistics: Data Mining and Remote Sensing" National Agricultural Statistics Service 3251 Old Lee Highway, Fairfax, VA 22030 - JSM 2009
- [13] Kant, Srivastava Uma. "Agro-Processing Industries: Potential, Constraints and Task Ahead". No. WP1989-10-01_00902. Indian Institute of Management Ahmedabad, Research and Publication Department, 1989.
- [14] Gandhi, Vasant, Gauri Kumar, and Robin Marsh. "Agroindustry for rural and small farmer development: issues and lessons from India", *The International Food and Agribusiness Management Review* 2, no. 3-4 (1999): 331-344.
- [15] Mollinga, Peter P. "The rational organisation of dissent: Boundary concepts, boundary objects and boundary settings in the interdisciplinary study of natural resources management". No. 33. ZEF working paper series, 2008.

Identifying Dynamic Topics of Interest across Social Networks

Mohamed Salaheldin Aly, Abeer Al Korany

Department of Computer Science
Faculty of Computers and Information
Cairo, Egypt

Abstract—Information propagation plays a significant role in online social networks, mining the latent information produced became crucial to understand how information is disseminated. It can be used for market prediction, rumor controlling, and opinion monitoring among other things. Thus, in this paper, an information dissemination model based on dynamic individual interest is proposed. The basic idea of this model is to extract effective topic of interest of each user overtime and identify the most relevant topics with respect to seed users. A set of experiments on real twitter dataset showed that the proposed dynamic prediction model which applies machine learning techniques outperformed traditional models that only rely on words extracted from tweets.

Keywords—Information propagation; topic modelling; dynamic user modelling; user behavior; machine learning; topic classification; social networks

I. INTRODUCTION

Billions of users are now using different social networks (SN), SNs have proven to be effective in communication. Understanding how information propagates across SNs became crucial to enhance the social networks and it attracted many businesses for the marketing value. Targeted advertisement along with many business applications in the past few years have proved to be very effective and to ensure the maximum efficiency researches have been studying information propagation in major social networks. This effort yields to develop different models that aim to predict how the information would propagate and its speed along with which users could be good candidates of becoming seeds for the information to propagate.

Information propagation depends on the users profiles which is represented by their interests, behaviour, and their position in the network which will affect their influence among other users. User's profiles contain a set of attributes that uniquely express each user like biography, age, gender, geographic location, hobbies, education history, and work information. While, other attributes that represent dynamic features with tagged time slots such as posts, comments and check-ins. Such information can be analysed in order to be used in different research areas such as: community detection, user recommendation (Abel et al., 2011; Blanco-Fernández et al., 2011). Studying user behaviour in SNs is quite complicated and the modelling for such behaviour has evolved drastically from how [Julia Stoyanovich 2008 et al.] [1] have simplified the user behaviour in their basic interests extracted from the

tags they frequently use in the URLs they publish. Understanding that SNs users' behaviour is dynamic requires the consideration of the temporal factors [2], [3] when categorizing the user behaviour. This paper proposes a dynamic user modelling framework that aims to predict the candidate seeds (set of most influencing users) in the social network who will be able to propagate information using topics of interest. The paper is organized as follows: Section II starts with discussing the related work. Section III introduces our dynamic user model, whereas Section IV explains the experimental setup used in building and validating that model. Section V discusses the results, and Section VI evaluates the model results. Section VII discusses the limitations while the paper is concluded in Section VIII.

II. RELATED WORK

Various traditional approaches have been proposed for information propagation. Popular topic models such as Latent Dirichlet Allocation (LDA) [4] assumed that users could be classified according to the tags extracted from the topics they share and their similarities. Given that the behavior of the SNs users is not static and that it changes over time, many efforts went into understanding the effect of the temporal factor over the extracted interests. Qiaozhu Mei and ChengXiang Zhai in 2005 [5] explored the temporal text mining by utilizing the timestamps from the social posts extracted to identify different patterns in the topics extracted over time, proposing that adding the temporal factors with the understanding of the nature of the topics propagating may explain the themes that might follow and how they could influence other topics. This is more obvious when it comes to news as with an event happening thousands of articles are written and posted, however after this sudden burst for that particular event rests the summary of such events are the ones that are propagated afterwards [6], therefore understanding the lifecycle of a thread is important. Xuerui Wang, Andrew McCallum [7] later on proposed A Non-Markov continuous-time model of topical trends where the extracted topics from a document could be considered as a constant yet that only constitutes the meaning of that particular document and that time is a variable that affects the correlation between the keywords in documents with similar topics afterwards.

In the above related work the focus is on the topic modelling and understanding the impact of the temporal factor on in the information propagation, yet in social networks information propagation is not only associated with topics.

Rather, user behavior as equally contributes to information propagation, especially that the behavior is not uniform. A Temporal Context-Aware Model for User Behavior was proposed by [8] which takes the two factors in consideration: 1) the users' interests 2) the temporal context in the topic selection. The model aimed for rating the nature of the user behavior (clicking, sharing, purchasing), it has an edge over previous studies as its able to differentiate between user oriented topics and temporal topics this enables the models to better understand the users' interests. The proposed model was tested on multiple social networks (Delicious, digg, movieLens and douban movies).

The Temporal Context-Aware Model was later on enhanced in [9] taking in consideration that users' interests across social networks are not stable yet the temporal factor has a huge impact on those interests. Given that a user's interests were capture at a point in time those interests will certainly change with changing his job, getting married or having a new born for example, hence users' interests are dynamic. The Dynamic Temporal context-Aware model considers the users' interests distribution across time to predict the likelihood of a user to interact with a social post at a certain point in time.

III. DYNAMIC MODELING OF USER

Information in social network is spread the interactions between different users or nodes. A node in a social network is an abstract representation of many features that identify it. Thus, users in a social network could be distinguished through several characteristics such as interests, behavior, activities, etc. Those characteristics are identified using either the content published by users or by analysis of their relationships through network links. Extracted content posted by the user is used to identify the user interest, while link-based features are used to identify the behavior and degree of influence between users. The proposed model utilize the content published by the users in order to predict the potential candidates to propagate specific content. As a case study, the proposed model was applied on Twitter dataset. The proposed model decomposes three main phases. The first phase aims to dynamically extract topic of interest of user. While, the second phase aims to classify users based on their topics of interest. The third phase identifies the topics to be spread by specific user within specific set of users by considering the effect of time. Each of those phases are described in the following subsections.

A. Extract User Dynamic Profile

The first phase in the proposed model is responsible for creating the dynamic user profile in terms of her/his topical interest. Thus, topics which represent interest of users within specific time interval are associated with their relevance (score). In micro blogging networks such as Twitter, the interests of a user could be extracted from two main sources: 1) the content that user publishes by her/himself, 2) the content that the user interacts with different neighboring in form of retweet and replies. Using both sources, interest of the user could be identified. It is significant to mention that the frequency of producing such content is also considered and used as a decaying factor to adjust the weights of the users' interests. As mentioned earlier, the proposed model differentiates between three types of topics of each users,

actual topics, burst topics, and pattern topics. Actual topics corresponds to frequent topics published by user represent user interest. When breaking news or events occur, people can post tweets about breaking news and share with friends, which could not be considered to represent a user interest. Due to large number of people participating in such conversation and discussion, those tweets may become hot messages and the source of burst topics. While the third type could also be observed where the content is triggered by an event yet the behavior is repeated every specific period of time such as Halloween. In order to be able to differentiate between each topics per user, topics of each user is extracted and associated with time slices, then if a topic is only mentioned in a specific time slice and then disappear from user topic list, then this could be categorized as a temporal topic . While pattern topics are extracted if it appear in the same period of each year.

B. Topic Classification

After identifying set of actual content that represent the user interest, MALLEET (MACHINE Learning for Language Toolkit) was applied to extract topics of interest with its associated relevance weight from each tweet. It uses a simple way to analyze unlabeled text, by defining a topic as a cluster of words with similar meanings and distinguish between uses of words with multiple meanings. A Java Wrapper was built to use MALLEET to analyze the collected tweets using Naïve Bayes algorithm and divide them into a set of topics to be pushed to IBM Watson to label it. For each topic i we calculated its relevance score with respect to the target user during time interval t as shown in (1):

$$TopRelScr_{(i,t)} = \frac{\sum_{j=0}^n OCRelWeight_{(i,j)}}{N} \quad (1)$$

Where n is the total number of occurrences of a topic i in tweets of the target user that are created within time interval t , N is the total number of topics contained in tweets of the target user within time interval t and $OCRelWeight$ is the relative weight provided by MALLEET of topic i for each one of its occurrence j in a tweet during time interval t . Finally, within each time interval t , each user's topical profile will be represented as a vector of topics associated with their relevance scores. It is significant to mention that, time plays an important role in calculating the topics relevance as well as the influence of a user. As time distribution of post behavior reflects massive users' behavior characteristic for burst and actual topics.

C. Identify Influence Users

Identifying potential "Influencers" over time is not an easy task, it is required to understand the position of each user in the network at a given time slice. Certain nodes that are established around specific topics are the seed to create the burst in social media. For example for football pages they share hundreds of posts during the match day The user's position in the network is defined by his influence which could be captured in a microblogging network such as twitter using different attributes that are available in the public dataset. The number of users following a certain user could be a simple way to indicate how influential he could be, the number of times his posts are favored or retweeted or the number of times he's mentioned in different users posts. To even measure such influence in a certain time slice we factor in the frequency per

tweet per time slice. However and as by definition of a network, the position of a certain users cannot be only determined by his behavior yet also the neighboring nodes in the network, for example a user can have a lot follower yet they would be information seekers with passive behavior and would not contribute to propagating the created content, yet on the other hand a user with fewer followers yet very active on the network could have much more influence. Thus we propose that the user's influence could be measured by how much other users interact with the content he shares along with the position of his followers in the network.

For each user we determine the following, the number of times his tweets were favored in a time slice equals the summation of all tweets favorite count over number of tweets in a time slice.

IV. EXPERIMENT SETUP

Twitter was selected as the social network to test the proposed model as it provides an easy to use API to extract data from public users. The API has its limitations yet enough data for testing purposes as it can provide all the tweets for a specific user during a specific time slice (with a limit of 3,000 tweets per user) along with the number of interactions on each tweet. The API also allows the retrieval of the list of friends for each user (with the limit of 5,000 per user) along with the total number of friends and followers.

A. Extracting Seed Users

For the purpose of this study a random sample of 1,000 public users was extracted using a java application to collect the data using twitter API and save them in an SQL relational database to facilitate the analysis. The sample was collected only from one location (Liverpool – UK) for two main reasons: 1) allow a better understanding of the context of the researched sample to facilitate the understanding of the contextual trends. 2) Understanding the influence of each node in the surrounding neighbors in the geographical network. The friends were also extracted for each user with the limitation of 5,000 users per user, accordingly 1,401,801 user where collected out of 3,884,033 in the 1000 users friends' lists. We selected the ones who were active during a specific time period which started from 1/1/2015 till 1/1/2016 regardless of their rate of tweet as our sample.

B. Extracting Users' Tweets

The 1,000 users collected had in total 12,793,079 Tweets. Given that the Twitter API has a limitation of around 3,000 tweets per user, only 2,248,181 tweets were collected. The tweet could be a retweet and accordingly the retweeted flag allows the differentiation between the content that is actually generated by the user and the content that the user shares from his network. Table I represent the summary of frequency of extracted tweets.

TABLE I. SUMMARY OF FREQUENCY OF EXTRACTED TWEETS

Tweet Type	Percentage
User Replies	15%
Tweet Replies	13%
Retweets	30%

Tweet Type	Percentage
Original Tweets	42%

C. Topic Modeling

One of the important challenges with the collected data is to be able to extract topics from the text for each tweet and differentiate them into corresponding types for each user. In a network like Twitter the issue becomes particularly complicated as the character limitation restricts users' accordingly they use abbreviations or slang that is difficult to classify. Understanding the content is not straight forward as for example in twitter the tweets are very short (with a maximum of 140 characters), accordingly even using different topic extractors such as Open Calais or IBM Watson the accuracy of the topics extracted is not reliable.

TABLE II. EXTRACTED TWEETS CATEGORIZATION

Category	Number of Tweets
Sports	541,516
art and entertainment	461,434
business and industrial	168,769
food and drink	158,400
law, govt and politics	122,273
Travel	101,655
Uncategorized	88,829
technology and computing	74,203
family and parenting	64,997
Society	62,130
Education	57,764
Shopping	55,835
Science	51,743
health and fitness	50,760
News	43,606
hobbies and interests	39,657
style and fashion	23,513
religion and spirituality	16,989
home and garden	15,985
Pets	15,450
Finance	15,245
automotive and vehicles	12,223
real estate	5,205
Grand Total	2,248,181

Using topic extractors is crucial to also allow the classification of topics extracted and understanding the areas of interest of each users on different levels. Thus, MALLET was used to analysis unlabeled text, and the Java Wrapper analyses the collected tweets and divide them into 500 topics, each topic having the top 50 significant keywords. The number of

iterations was set to 2000 to refine the results as much as possible, 19,272,829 tokens were found in all the tweets collected and used to train the model and create the 500 different topics. After training the model it was then used to go through all the tweets and assign each tweet to the most relevant topic with a relevancy score. Since the 500 topics were in the form of a cluster of related keywords yet not labelled, each cluster was then pushed to IBM Watson to label it. Watson API offers different configuration settings to get the desired output, for the purpose of our experiment the categories, entities and concepts were selected each with a set limit of three. Watson was only able to Identify 468 topics out of the 500 giving an output of 154 category for all the tweets collected. The summary of the categorized tweets is shown in Table II forming 20 categories.

V. TOPIC CLASSIFICATION

A. Extracting Bursty Topics

We take an example from the collected dataset to better understand the differentiation between the temporal topics and the topics that are based on user interest. Since the data collected is only in Liverpool, we take the hashtag “Cunard175” where Cunard liner a Britannia ship, left British waters bound for America marked its 175th anniversary in Liverpool, the event was on May 2015 by looking at the normal distribution of the topic over 2015 in Fig. 1 we find the following:

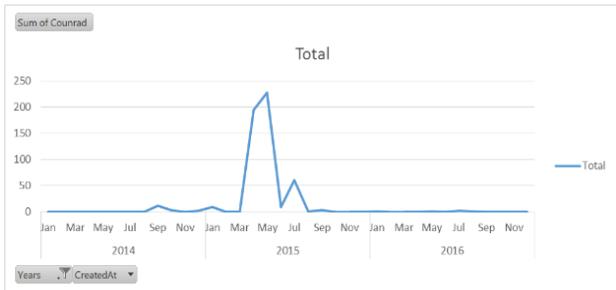


Fig. 1. Normal Distribution of the Topic Over 2015.

The topic was purely generated by temporal trigger in April and May 2015. Thus, it could not be considered as one of the interests of any of the users who has shared it. Thus, for each year, burst topics were detected and eliminated from all users’ topics of interest using the following algorithm.

If number of tweets for topic is greater than twice the calculated median mark as burst topic Another example would be the sudden burst in the “/law, govt and politics/law enforcement/police” interest, in Table III we can see the significant score in April 2016.

TABLE III. SIGNIFICANT SCORE OF DIFFERENT TOPICS IN APRIL 2016

Topic	Median	Score	YYYY-MM
/law, govt and politics/law enforcement/police	57	2.56	2016-01
/law, govt and politics/law enforcement/police	57	2.28	2016-02
/law, govt and politics/law enforcement/police	57	2.46	2016-03
/law, govt and politics/law enforcement/police	57	9.05	2016-04

To understand the reason for the burst we start looking into major events or news that are relevant to the topic identified and explore different possibilities. For example, Hillsborough disaster which was a human crush at Hillsborough football stadium in Sheffield, England on 15 April 1989, during the 1988–89 FA Cup semi-final game between Liverpool and Nottingham Forest. The resulting 96 fatalities and 766 injuries makes this the worst disaster in British sporting history which came shortly after the 27th anniversary of the lethal crush at the FA Cup semi-final between Liverpool and Nottingham Forest, vindicated the bereaved families. The number of tweets increased started increasing from January until it reached 9 times its median in April 2016. Similarly thirty nine topics were identified to have bursts throughout 2016.

B. Extracting Pattern Topics

The second type could also be observed where the content is triggered by an event yet the behavior is repeated every specific period of time. We take Halloween as example where and check the normal distribution over four years of data, we notice that every year around October there is a spike in the number of mentions for this topic as shown in Fig. 2.

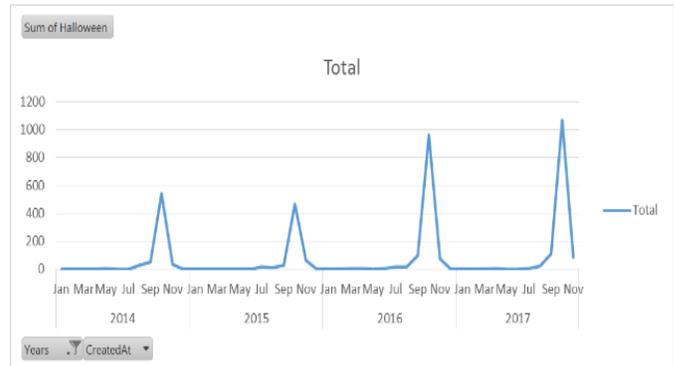


Fig. 2. Example of Pattern Topics (Halloween).

C. Extracting Actual Topics of Interest

Finally, the remaining topics of each user are considered her/his topic of interest. For example, Liverpool FC, this topic is constantly mentioned by different users over time and is not a temporal topic although bursts could be observed in some time slices, however those bursts could be attributed to certain contests in the context of Liverpool FC as shown in Fig. 3.

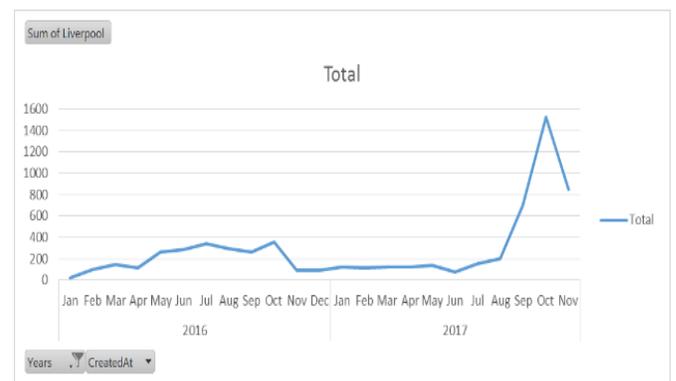


Fig. 3. Example of Actual Topics of Interest.

D. Refining the Sample

To be able to test the proposed model the data collected had to be refined to ensure that for each user, only actual topic of interest would be used in prediction. Thus we identify active users that had tweets during 2015, 2016 and 2017. The proposed model would be applied to extract user interests on 2015 and 2016 and use the results to feed in the overall model and run it on 2017 for evaluation. Accordingly we choose 631 users having tweets in all three years and we start detecting their interests by first categorizing the different tweets to see where they fall in the three categories mentioned above.

E. Identifying Dynamic User Interests

Fig. 4 shows the change of percentage of each interest from the overall interests of one user across time. For example, “Shopping and gifts” had 73% of the overall interests and further explore as the profile is for a famous footballer where the algorithm was responsive for the event accordingly gave it a high percentage among the interests while the month after the curve had a dive equivalent to the hike it had in February 2014. As a result the topic would not have a score increase yet and thus, decay factor should be considered over time. For example in Fig. 4, the hike in February 2014 affects the percentage the topic has in the user’s interests despite the fact that the interaction with the topic for slices after was minimum.



Fig. 4. Change of One Topic of Interest of One User Over Four Year Without Considering the Decay Factor.

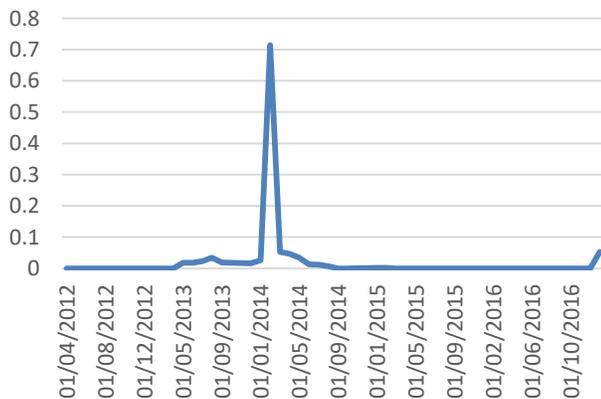


Fig. 5. Change of One Topic of Interest of Same User Over Four Year Using a Decay Factor.

While when considering the decay factor, the topic appears in slice yet with a lower score as shown in Fig. 5.

$$\text{New score} = \frac{\text{currentScore} - \text{tweetsInPreviousSlice}}{\text{TweetsInSlice}} \times \frac{(\text{tweetsInPreviousSlice} + 1)}{2}$$

The above method ensures the reversal of any burst effect for any topic by calculating the ratio of tweets in the current slice to the tweets in the one preceding it for the same topic and using it as a multiplier to half the number of tweets in the previous slice.

VI. EVALUATION

We then evaluate the results of the model by calculating the accuracy of the predicted number of tweets per topic for each user by applying the following for each tweets in 2016 and January 2017:

- For each user use the model score once with the decay factor and once without, to predict the number of tweets per topic for January 2017.
- Collect the actual tweets over January 2017 for the same users and use MALLET and IBM Watson to categorize them per topic
- Compare the actual number of tweets per category published for each user with both predicted scores by calculating the accuracy score for each.

The results show that the without the decay factor the model is 46% accurate while after applying the decay factor the model becomes more accurate as expected with 60% accuracy.

VII. LIMITATIONS

The public data that could be extracted from twitter for modeling is limited not only when it comes to quantity but also the behavioral data that could be crucial for this research such as the tweets favored or retweeted by each user. Such data could significantly enhance the model by factoring in those attributes in the weighting process.

VIII. CONCLUSION

The dynamic behavior of users across social networks makes it extremely challenging to predict user interests and perfectly understand how information propagates across social networks, However it is possible to reduce the factors that might decrease the accuracy of the predictions which we tried to do in this paper by understanding the nature of the interests of each users and eliminating all behavior that is not considered steady enough to predict future tweets. With this understanding and with a flexible design for the predictive model it is possible to use machine learning to profile users using their different activities and enhance their experience on social networks by displaying the most relevant content along with the utilization of the marketing value.

REFERENCES

[1] J. Stoyanovich, S. Amer-Yahia, C. Marlow, and C. Yu. A Study of the Benefit of Leveraging Tagging Behavior to Model Users’ Interests in

- del.icio.us. In AAAI Spring Symposium on Social Information Processing, 2008.
- [2] Abel, F., Gao, Q., Houben, G.-j. and Tao, K. (2011) 'Analyzing User Modeling on Twitter for Personalized News Recommendations', in *User Modeling, Adaption and Personalization*, pp.1-12.
- [3] Blanco-Fernández, Y., López-Nores, M., Pazos-Arias, J.J. and García-Duque, J. (2011) 'An improvement for semantics-based recommender systems grounded on attaching temporal information to ontologies and user profiles', *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 1385-1397.
- [4] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. *J. Machine Learn. Res.* 3, 993–1022
- [5] Qiaozhu Mei and ChengXiang Zhai Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining. ACM 1-59593-135-X/05/0008
- [6] Qiming Diao, Jing Jiang, Feida Zhu, Ee-Peng Lim. Finding Bursty Topics from Microblogs. ACL '12 Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Volume 1 Pages 536-544
- [7] Xuerui Wang, Andrew McCallum 2006 Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. ACM SIGKDD-2006 August 20-23, 2005, Philadelphia, Pennsylvania, USA
- [8] Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, Zi Huang. Temporal Context-Aware Model for User Behavior. Proceeding SIGMOD '14 Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data Pages 1543-1554
- [9] Jia Dong Zhang and Chi Yin Chow. Ticrec: A probabilistic framework to utilize temporal influence correlations for time-aware location recommendations. *IEEE Transactions on Services Computing*, (1):1–1, 2015.

Processing Sampled Big Data

Waleed Albattah, Rehan Ullah Khan
Information Technology Department
Qassim University,
Qassim, KSA

Abstract—Big data processing requires extremely powerful and large computing setup. This puts bottleneck not only on processing infrastructure but also many researchers don't get the freedom to analyze large datasets. This paper thus analyzes the processing of the large amount of data from machine learnt models that are built on the smaller sets of data samples. This work analyzes more than 40 GB data by testing different strategies of reducing the processed data without losing and compromising on the detection and model learning in machine learning. Many alternatives are analyzed and it is observed that 50% reduction does not drastically harm the machine learning model performance. On average, in SVM only 3.6%, and in Random Forest, only 1.8% performance is reduced, if only 50% data is used. The 50% reduction in instances means that in most cases, the data will fit in the RAM and the processing times will be considerably reduced, benefitting in execution times and or resources. From the incremental training and testing experiments, it is found that in special cases, smaller sub-sampled data can be used for model generation in machine learning problems. This is useful in cases, where there are either limitations on hardware or one has to select among many available machine learning algorithms.

Keywords—Deep learning; content analysis; machine learning; support vector machines; random forest

I. INTRODUCTION

The immense increase in technology sophistication these days and the relevant exponential increase of data being circulated and produced has resulted into the fact that ordinary data has been turned into big data. As explained by the name, big data refers to the data type that is massive in its size, formats that it holds, and requires high velocity servers for fetching it at required time [1]. Big data constitutes of variability, volume and velocity of data that needs to be accessed. This data is usually stored in large servers and is accessed only when required [2]. This big data is then used for carrying out ordinary operations of organization like decision making, sorting and other business related tasks [3]. However, for increasing efficiency and accuracy, a tradeoff between efficiency and size of application is crucial [4]. Common example of this is global positioning system, facial recognition cameras and connected automated vehicles. The efficiency of these applications can be enhanced through the provision of increased data sets for model learning. On the other hand, this is not feasible as large data sets require high storage space which in turn, becomes hard to be processed. For this purpose, it is required that a mechanism is built that allows sub sets of big data to hold similar knowledge and information as that of original data [5].

Big data has posed some serious risks to the data computation as well which needs to be addressed in order to ensure that the end user is protected in the end. For this purpose, usually some parameters are defined which ensure big data quality and information quality [6]. These parameters include, Syntactical Validity, Appropriate Identity association, appropriate attribute association, accuracy, precision, temporal applicability, theoretical relevancy, practical relevancy, currency, completeness, controls and audibility [6]. Other than this, management of servers and data for access control, privileges, sortation and security create other issues [7]. By 2002, digital devices were more than 92% with 5 Exabyte of data [8]. This number has been increasing since then and the problem has been evolving gradually. Today, big data is about \$46.4 billion industry [8], meaning that, despite the problems of data handling, interest of users is growing over years. When it comes to data mining, this task becomes extremely complex when there exist hundreds of groups that are classified on the basis of minor differences, increasing work load and compilation time [8].

Apart from its never ending applications, big data is becoming a challenging concept for data mining, machine learning, information fusion, computational intelligence, social networks and the semantic web, etc. [9]. In this regards, issues of data processing, data use for pattern mining, data storage, user behavior analysis, data visualization and data tracking have attracted considerable attention [10].

This havoc of solution search for big assembly issues has been increased due to the fact that technologies like machine learning, computational intelligence and social networks are using libraries for data processing. These libraries are in turn increasing in size as the application scope is increasing. Due to which, solutions for simplicity of big data handling are continually researched and examined. These solutions include, data sampling, data condensation, density based approaches, incremental learning, divide and conquer, grid based approaches, distributed computing and others [8].

From processing perspectives, the Big data sampling has the biggest issue of complexity, computational burden and inefficiency to complete the task properly [11]. Sampling effort is the number of data sets that can be added per sample. It is assumed generally that sampling effort data sets richness is weak only if sampling bias is done successfully through estimation [12]. Selection bias, in this regards, can be computed and determined successfully though inverse sampling procedure, in which information from external resources is used, or by digital integration technique, in which big data is combined with independent probability sample

[13]. Size of a sample is extremely critical and plays an important role in determining accuracy of the system [14]. For this purpose, as a solution to big data sampling issues, many algorithms have been presented like Zig Zag process [15], non-probability sampling [13], inverse sampling, cluster based sampling [16].

Machine learning is a part of data analytics that learns from the available data to predict, decide and take insights [16]. Based upon statistics, it extracts trends from data and then computes it for supervised or unsupervised learning techniques. In machine learning, machines are made to understand information and made capable to derive some meaning out of them. This learning is done through analogies, connectionist, strategies, discovery, problem solving, search, and match by parameter adjustment. The ability of any machine to learn depends upon the amount of information it can handle and the limit to which it can process [4]. Machine learning is considered as the type of automation that gets enhanced as the amount of input data increases. However, algorithms being used for computation are usually conventional that are designed to solve simple data sets, creating a computational challenges. For example, for big data these are memory and processing for training periods, unstructured data formats, fast moving data, low scalability of algorithms, unbalanced distribution of input data sets, and unlabeled data [17].

Convolutional Neural Networks (CNNs) have been used for accurate modeling of classification data [18], [41], especially image and text data [42]. However, for large datasets, the CNNs needs tremendous amount of processing power. Also, the trend has moved from the traditional feature extraction to autonomous feature extraction as in [19], [20]. However, the main problem is still not thoroughly investigated, which is the increase of features and data instances lead to the curse of dimensionality and the tremendous amount of processing power needed. The curse of dimensionality definitely affects the final model performance. Similarly the continuous increase of data instances forces the machine learning models to be re-calculated and re-evaluated. This thus puts tremendous reliance on the powerful computing machines and resources. However, such facilities are still not available to masses and many research institutes.

This article thus investigates the reduction of data instances for classification performance and machine learning scenarios. For experimental evaluation of the proposed architecture, this article uses the dataset from the NDPI videos. Further details are available in [21]. NDPI is huge dataset and comprises of more than 40 Gigabytes of videos data. For experimental analysis, the data is divided into three classes. These are: Un-acceptable, Acceptable, and Flagged. Though the paper is based on the generic concept of data sampling and performance analysis, however, the article uses the data from image based filtering. It has three main reasons. First is that the data is well organized into three classes, which is a good representative problem for machine learning algorithms? Secondly, though the data is image, in the feature form, the data is converted to numerical values. Thus the data is equated to other datasets and similar machine learning problems. Thirdly, the data is huge, more than 40 Gigabytes in

size. Therefore, it is assumed that the data that is processed in this article is big data. Therefore, the results can be extended to other datasets of similar nature.

Our previous work [1] about the role of sampling in big data analysis motivated us for further investigation about effective approaches for big data analysis. Based on the dataset processed in this article, there is considerable work available in the state of the art. The articles [22]-[25] present and models such scenarios and applications.. The work in [22] fuses AlexNet [20] and GoogLeNet [26] and for performance enhancements. The work in [24] takes advantage of colors transformations. The paper [27] presents an evidence combination. The work of [28] takes advantage of adaptive sampling approach for filtering. The paper [29] demonstrates websites filtering analysis, and [30] combines key-frame analysis. The [31] and [32] use visual features for media access and filtering. The articles [33]-[36] are based on content based image retrieval.

The rest of the paper is organized as follow. Section II presents some background about the classifiers used for the study, namely, Support Vector Machines and Random Forest. Section III explains the experimental study and the found results. Discussion of results is presented in Section IV. Finally, Section V concludes the study.

II. CLASSIFICATION

Classifiers draw a decision boundary between the classes in the data. There are several classifiers present. In this article, we use the Support Vector Machine (SVM) and the Random Forest. The SVM has shown great results and in a favorable choice in machine learning and computer vision tasks. Moreover, SVM has been heavily used with the DL features learning and training. The Random forest has also shown considerable good results and is the choice in many classification scenarios.

A. SVM

Support Vector Machine (SVM) is a supervised learning classifier that is introduced in 1990s by Boser, Guyon, and Vapnik [37]. It is widely used because of its accuracy, ability to deal with high-dimensional data, and its flexibility in modeling different sources of data. The SVM has two advantages: first, it has the ability to produce non-linear decision boundaries by using methods of linear classifiers; secondly, the classifier can be applied to data with no fixed-dimensional vector space representation [38]. Moreover, SVM has a robust theoretical foundation, which is statistical learning theory; and successful empirical applications as well. It has been applied to different fields such as hand written digits recognition, text classification, and objects recognition [38]. The SVM in this article is used due to its over-all good detection performance in similar areas.

B. Random Forest

Two popular methods of classification trees have grabbed researchers' attention: bagging and boosting. These two methods can generate many classifiers and aggregate their results [39]. One of the important advantages of Random forest is that it can be used for regression or classification

problems. In an enhancement addition to bagging, Breiman [40] proposed random forests as an additional layer of randomness. Either in regression or classification problems, Random forest can help in ranking the importance of variables. Random forest has only two parameters: the number of trees in the forest and the number of variables in the node. These two parameters constitute to the straightforwardness of Random forest. Moreover, it constructs every tree with a different bootstrap sample of data, which changes how trees are constructed in regression and classification. Each node is split by the best predictor chosen at the node randomly among a subset of predictors [40]. Many trees are grown and every tree vote for a particular class. The class with high number of trees is the final class assigned to particular data instance.

III. EXPERIMENTAL SETUP AND RESULTS

A. Features and Dataset

For an evaluation experiments, the article uses datasets from the NDPI videos. Further details of the NDPI dataset is available in [21]. NDPI is huge dataset and comprises of more than 40 Gigabytes of videos data. For experimental analysis, the data is divided into three classes. These are: Unacceptable, Acceptable, and Flagged. Fig. 1 shows some samples. The experiment setup uses the data from image based filtering and large amount of data. It has three main reasons. First is that the data is well organized into three classes, which is a good representative problem for machine learning algorithms? Secondly, though the data is image, in the feature form, the data is converted to numerical values. Thus the data is equated to other datasets and similar machine learning problems. Thirdly, the data is huge, more than 40 Gigabytes in size. Therefore, it is assumed that the data that is processed in this article is big data. Therefore, the results can be extended to other datasets of similar nature.

For feature extraction, the article uses the Autocorrelogram. We use the F-measure as an evaluation parameter as it is mostly used in the state of the art for similar problems and applications and is favorable for this evaluation as well. The F-measure takes into account the Precision and the Recall.



Fig. 1. Sample Images from NDPI [21].

B. Instance Sampling

In data analysis domains, an instance represents the individual object of which the problem is composed of. This means that if the problem is based on the color, let's say in computer vision, the instance is the set of pixels for the problem concerned. The instance may also represent a complete image if the features are globally extracted from the images. In most cases, the instance is directly related to the

number of objects available for training and testing. If instances are reduced, the training data and ultimately testing data is reduced. If instances are increased, it will mean that the training and testing data is increased. If a ten folds cross validation is used, instance increase results in the increase of 90% training samples, and 10% testing samples. This can have one of three impacts on the results. The result could stay neutral. It can increase in certain cases, and it can also decrease in certain cases.

The neutral case can occur generally in two ways. First one is if the instance added has similar nature to the previous data. This means that the instance is already represented in the model of the machine learning classifier. The addition of this new instance thus has either contributed no extra information. This thus increases the dataset without any benefit to the machine learning model. The second neutral case is when contribution of the instance addition is negligible due to the large number of data samples. This can also mean that the data is already covering most of the model generation cases and no extra addition of data is required.

The increase in classification results due to instance increase can be due to the fact that the new instances contribute strong classification information in the model. It means that the new addition strongly represents the classes in the dataset and also exhibiting strong correlation with the attributes for that instance. This type of scenario is always the objective in machine learning paradigm. However, every machine learning algorithm has certain limits and adding more strong instances may not contribute any information for classification. One of the interesting phenomenon that can occur by adding strong instances is over-fitting. The model can become much diverted to special cases and does not generalize well.

The decrease in performance can be due to either the new instances are not related to the classes in specific problem, or the instance added is representing (adding) strong noise to the model. This phenomenon is most common and collecting correct dataset is the challenge for most machine learning related problems. Therefore, the data cleaning task is essential in many classification tasks for reliable model generation. The decrease in classification performance can also be due to less number of data instances. Many machine learning algorithms require considerable amount of data (not big data) for reliable model generation and generalization for unseen test data instances. However, this does not mean that data increase beyond certain limit will keep on increasing the performance. Every classifier has limits for certain problems and thus thorough analysis is required for the final model generation and the amount of data needed for the particular problem.

In the experimentation setup, the objective is to analyze instance addition and removal on the results of machine learning. The experimental approach however proceeds in reverse manner. The proposed experimentation setup proceeds by reducing instances and analyzing the results. The 50% instances are sampled from original 100% data and the results are noted. The 50% sample is randomly selected from the original data. The sample is thus analyzed based on 90% training data and 10% testing data. This means that from

particular 50% data, the 90% of the data is used for training the classifier for model generation. This model generated thus is used to test the 10% data of the 50% sample and performance is noted. This 90% training data the 10% training data from the 50% of the original data is also selected ten times to take average 10 ten permutations. This is to remove biasness. This of taking 50% training data is then repeated 100 times. This generates 1000 experiments.

Fig. 2 shows experiments where the 50% instances are randomly selected for the actual 100% data. Fig. 2 shows performance in terms of an F-measure for the 3 classes' data based on the SVM classifier. The "Actual Data" label in Fig. 2 represents the F-measure of the original 100% data. From Fig. 2, an F-measure of 0.784 is obtained for the 100% data. The F-measure for the average of first ten experiments (labelled "10" in Fig. 2) is 0.733 which is less than the F-measure of the 100% instances. The F-measure for the average of next ten experiments (represented as "20" in Fig. 2) is 0.703. The third set has an F-measure of 0.755. The fourth, fifth, and sixth set has an F-measure of 0.762, 0.757, and 0.77, respectively. The seventh set has a reduced F-measure of 0.722. The eighth and ninth sets have an F-measure of 0.744 and 0.741. The last set of experiments gets an increased F-measure of 0.797. This is even higher than the F-measure of the 100% original set. The average of the all 100 experiments (labelled as "Average of 100") is 0.748. As the F-measure for 100% data is 0.784, therefore, the difference is 0.036. This means the total difference of 3.6% to the original 100% data. This thus means that the model of 100% data is 3.6% more accurate than the sampled data.

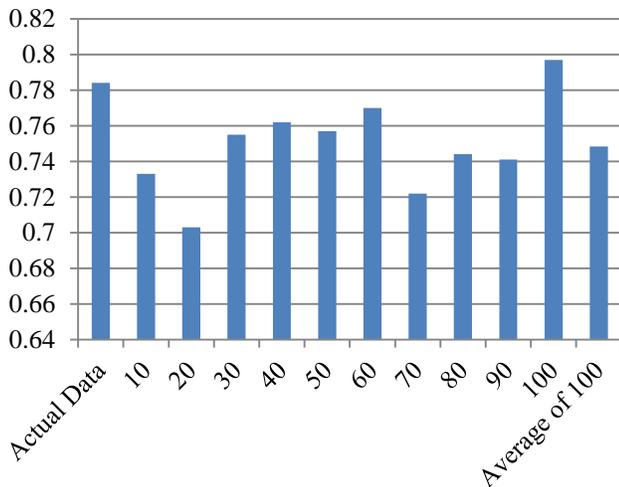


Fig. 2. F-measure for the 3 Classes' Data based on the SVM Classifier for 50% Random Instances. The 10 Interval Sets Show the Average F-measure of 10 Experiments.

Fig. 3 shows experiments with the Random Forest and almost follows on average the SVM scenario. The "Actual Data" label in Fig. 3 represents the F-measure of the original 100% data. From Fig. 3, an F-measure of 0.841 is obtained for the 100% data. The F-measure for the average of first ten experiments (labelled "10") is 0.83 which is less than the F-measure of the 100% instances as was with the case of SVM. The F-measure for the average of next ten experiments

(represented as "20") is 0.809. The third set has an F-measure of 0.822. The fourth, fifth, and sixth set has an F-measure of 0.836, 0.824, and 0.814 respectively. The seventh set has a reduced F-measure of 0.818. The eighth and ninth sets have an F-measure of 0.84 and 0.816. The last set of experiments gets an F-measure of 0.83. The average of the all 100 experiments (labelled as "Average of 100") is 0.823. As the F-measure for 100% data is 0.841, therefore, the difference is 0.018. This means the total difference of 1.8% to the original 100% data. This thus means that the model of 100% data is 1.8% more accurate than the sampled data.

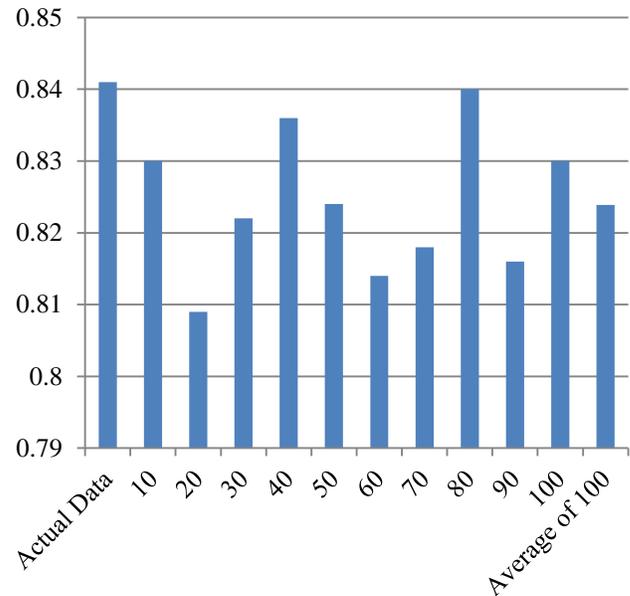


Fig. 3. Random Forest Classifier F-measure for the 3 Classes' Data based on the 50% Random Instances. The 10 Interval Sets Show the Average F-measure of 10 Experiments.

C. Incremental Training and Test Data

The instances of data can also be analyzed based on the amount of training and testing data available. Generally, the more training data, the better is the performance. This may be true in most problems; however, this may not be true in every case because the more the data, the more is the chance of noisy data and thus wrong models. Therefore, to analyze this, in the following set of experiments, the objective is to see if the impact of the amount of training data affects the performance of machine learning algorithms and classifiers. If a smaller set of data can generate good model that can generalize well and have good classification performance, then large of data may not be needed to process, thus saving time and computing resources. Moreover, this enables researchers to quickly analyze datasets on many algorithms which is useful for research and development activities.

Table I shows the F-measure based analysis of the SVM classifier with the incremental increase of training data. In Table I, Training data of "10" means that the 10% of the data is used for generating the model of the Random Forest classifier and 90% data is used for testing this model. The F-measure for such scenario is 0.648. Increasing the training

data to 20% and reducing the testing data to 80% gets an F-measure of 0.718. At 30% training data, the F-measure increases to 0.742. With 40, 50, and 60 percent, the F-measure obtained is 0.744, 0.776, and 0.745 respectively. At 70%, an increase F-measure of 0.768 is obtained. At 80%, the F-measure keeps increasing to 0.799. At 90% training data and 10% testing data, the F-measure normalizes at 0.806.

Table I shows that the increase in training data and its relation to the F-measure is not consistent in all cases. For example, the F-measure at the 60% training is lower than the 50% training, which in theory should be higher. Similarly, in case of the 70% training, the F-measure is less than 50% training.

TABLE I. F-MEASURE OF SVM FOR INCREMENTALLY INCREASING (BY 10%) THE TRAINING DATA STARTING FROM 10%. SIMILARLY, INCREMENTALLY REDUCING (BY 10%) THE TESTING DATA STARTING FROM 90%

Training Data	Testing Data	F-measure (SVM)
10	90	0.648
20	80	0.718
30	70	0.742
40	60	0.744
50	50	0.776
60	40	0.745
70	30	0.768
80	20	0.799
90	10	0.806

Table II shows the F-measure based analysis of the Random Forest classifier with incremental increase of training data. The F-measure 10% training data is 0.706. Increasing the training data to 20% and reducing the testing data to 80% gets an F-measure of 0.77. At 30% training data, the F-measure increases to 0.785. With 40, 50, and 60 percent, the F-measure obtained is 0.769, 0.806, and 0.803 respectively. At 70%, an increase F-measure of 0.854 is obtained. At 80%, the F-measure keeps increasing to 0.865. At 90% training data and 10% testing data, the F-measure normalizes at 0.854.

Table II shows the increase in training data and its relation with the F-measure is not consistent in all cases. For example, the F-measure at the 40% training is lower than the 30% training. Similarly, in case of the 90% training, the F-measure is less than 80% training.

Both in the Tables I and II, the increase in training data and its relation to the F-measure is not consistent in all cases. This could be due to many reasons. One of the reasons is that the increasing number of samples can add noise and thus more training data does not mean good final trained model. Secondly, since the selection of training data is random, the training sample does not pick many instances of the “good” representative samples.

TABLE II. F-MEASURE OF RANDOM FOREST FOR INCREMENTALLY INCREASING (BY 10%) THE TRAINING DATA STARTING FROM 10%. SIMILARLY, INCREMENTALLY REDUCING (BY 10%) THE TESTING DATA STARTING FROM 90%

Training Data	Testing Data	F-measure (Random Forest)
10	90	0.706
20	80	0.77
30	70	0.785
40	60	0.769
50	50	0.806
60	40	0.803
70	30	0.854
80	20	0.865
90	10	0.854

IV. DISCUSSION OF RESULTS

Experiments in both the Fig. 2 and 3 depict interesting results. With these experiments, it is observed that 50% reduction does not drastically harm the overall model. On average, in SVM only 3.6%, and in Random Forest, only 1.8% performance is reduced if only 50% data is used. This is acceptable in most cases unless there is a serious nature of the problem in hand. The benefit one gets is the processing of extremely reduced size and sets of data instances. This is useful in number of scenarios. 50% reduction in instances means that in most cases, the data will fit easily in the RAM and the processing times will be considerably reduced, benefitting in terms or resources. Other benefits are that since data generation and gathering is not an easy task, the less number of instances means that less but clean data can be useful in many cases.

Fig. 4 shows the follow of the F-measure plotted against the amount of training data. The X-Axis shows the training samples. The Y-Axis shows the F-measure. Fig. 4 shows slightly incremental increase in F-measure for both the SVM and the Random Forest in many cases. However, interestingly, it can be seen that even with the 10% training data, there is not a huge jump and difference in the F-measure of consecutive incremental sets of data in both the SVM and the Random Forest cases. From this it can be deduced that in special cases, smaller sub-sampled data can be used for model generation in machine learning problems. This is useful in cases, where there are either limitations on hardware or one has to select among many available machine learning algorithms. The second being the most common scenario. The first case of hardware resources is more defined in the case of processing big datasets. The ever increasing data has put tremendous limitations on the processing power of many available machines. Many researchers need special hardware to process large amount of data that is expensive and is mostly still not available to some research groups and teaching environments. This experimental setup shows that in special cases, generating many random models from the smaller samples and averaging its performance can represent larges datasets. This type of reliance on smaller models is thus useful in quick model analysis and experimentation, where the final model can be then generated by processing big datasets.

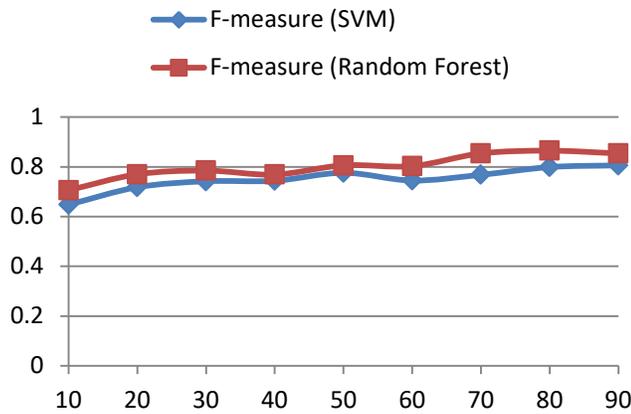


Fig. 4. F-measure for Incrementally Increasing (by 10%) the Training Data Starting from 10%. Similarly, Incrementally Reducing (by 10%) the Testing Data Starting from 90%.

V. CONCLUSION

Big data processing requires large computing resources. This puts bottleneck not only on processing data but also many researchers don't get the freedom to analyze large datasets. This article analyzed large amount of data from different perspectives. One of them is the processing reduced sets of large data with less computing resources. Thus the article analyzed 40 GB data, by testing different strategies of reducing the processed data without losing and compromising on the detection and model learning in machine learning. Many alternatives were analyzed and it is observed that 50% reduction does not drastically harm the machine learning model performance. On average, in SVM only 3.6%, and in Random Forest, only 1.8% performance is reduced if only 50% data is used. This is acceptable in most cases unless there is a serious nature of the problem in hand. The benefit one gets is the ability and freedom of processing of extremely reduced size and sets of data instances. This is useful in number of scenarios. The 50% reduction in instances means that in most cases, the data will fit easily in the RAM and the processing times will be considerably reduced, benefitting in execution, time and or resources. From the incremental training and testing experiments, it is found that in special cases, smaller sub-sampled data can be used for model generation in machine learning problems. This is useful in cases, where there are either limitations on hardware or one has to select among many available machine learning algorithms. The second point being the most common scenario in machine learning research. In future, the experimentation setup will be expended to massive parallel architecture for large collection of data sets including textual data. Also, the DL will be analyzed for sampled based training and testing.

REFERENCES

- [1] W. Albattah, "The Role of Sampling in Big Data Analysis," in Proceedings of the International Conference on Big Data and Advanced Wireless Technologies - BDAW '16, 2016, pp. 1-5.
- [2] M. Hilbert, "Big Data for Development: A Review of Promises and Challenges," *Dev. Policy Rev.*, vol. 34, no. 1, pp. 135-174, Jan. 2016.
- [3] D. A. Reed and J. Dongarra, "Exascale computing and big data," *Commun. ACM*, vol. 58, no. 7, pp. 56-68, 2015.
- [4] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz,

- "Machine Learning With Big Data: Challenges and Approaches," *IEEE Access*, vol. 5, no. 1, pp. 7776-7797, 2017.
- [5] K. Singh, S. C. Guntuku, A. Thakur, and C. Hota, "Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests," *Inf. Sci. (Ny)*, vol. 278, pp. 488-497, 2014.
- [6] R. Clarke, "Big data, big risks," *Inf. Syst. J.*, vol. 26, no. 1, pp. 77-90, Jan. 2016.
- [7] D. Sullivan, "Introduction to big data security analytics in the enterprise." [Online]. Available: <https://searchsecurity.techtarget.com/feature/Introduction-to-big-data-security-analytics-in-the-enterprise>. [Accessed: 31-Jul-2018].
- [8] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, vol. 2, no. 1, p. 21, Dec. 2015.
- [9] G. Bello-Ortiz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45-59, Mar. 2016.
- [10] J. Zakir, T. Seymour, and K. Berg, "Big Data Analytics," *Issues Inf. Syst.*, vol. 16, no. 2, pp. 81-90, 2015.
- [11] U. Sivaraman, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263-286, Jan. 2017.
- [12] K. Engemann et al., "Limited sampling hampers 'big data' estimation of species richness in a tropical biodiversity hotspot," *Ecol. Evol.*, vol. 5, no. 3, pp. 807-820, 2015.
- [13] J. K. Kim and Z. Wang, "Sampling techniques for big data analysis in finite population inference," Jan. 2018.
- [14] S. Liu, R. She, and P. Fan, "How Many Samples Required in Big Data Collection: A Differential Message Importance Measure," Jan. 2018.
- [15] J. Bierkens, P. Fearnhead, and G. Roberts, "The Zig-Zag Process and Super-Efficient Sampling for Bayesian Analysis of Big Data," Jul. 2016.
- [16] J. Zhao, J. Sun, Y. Zhai, Y. Ding, C. Wu, and M. Hu, "A Novel Clustering-Based Sampling Approach for Minimum Sample Set in Big Data Environment," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 2, pp. 1-10, Feb. 2018.
- [17] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, no. 1, pp. 350-361, 2017.
- [18] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, "From Group to Individual Labels Using Deep Features," *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '15*, pp. 597-606, 2015.
- [19] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915-1929, Aug. 2013.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc., pp. 1097-1105, 2012.
- [21] "Pornography Database." [Online]. Available: <https://sites.google.com/site/pornographydatabase/>. [Accessed: 09-Nov-2017].
- [22] M. Moustafa, "Applying deep learning to classify pornographic images and videos," Nov. 2015.
- [23] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, M. de M. Coelho, and A. de A. Araújo, "Nude Detection in Video Using Bag-of-Visual-Features," in 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, 2009, pp. 224-231.
- [24] A. Abadpour and S. Kasaei, "Pixel-Based Skin Detection for Pornography Filtering," *Iran. J. Electr. Electron. Eng.*, vol. 1, no. 3, pp. 21-41, 2005.
- [25] R. Ullah and A. Alkhalifah, "Media Content Access: Image-based Filtering," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 3, 2018.
- [26] C. Szegedy et al., "Going Deeper with Convolutions," Sep. 2014.
- [27] E. Valle, S. Avila, F. Souza, M. Coelho, and A. de A. Araujo, "Content-Based Filtering for Video Sharing Social Networks," in XII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais—SBSEG 2012, 2011, p. 28.
- [28] P. Monteiro, S. Eleuterio, M. De, and C. Polastro, "An adaptive

- sampling strategy for automatic detection of child pornographic videos.”
- [29] N. Agarwal, H. Liu, and J. Zhang, “Blocking objectionable web content by leveraging multiple information sources,” *ACM SIGKDD Explor. Newsl.*, vol. 8, no. 1, pp. 17–26, Jun. 2006.
- [30] C. Jansohn, A. Ulges, and T. M. Breuel, “Detecting pornographic video content by combining image features with motion information,” in *Proceedings of the seventeen ACM international conference on Multimedia - MM '09*, 2009, p. 601.
- [31] J.-H. Wang, H.-C. Chang, M.-J. Lee, and Y.-M. Shaw, “Classifying Peer-to-Peer File Transfers for Objectionable Content Filtering Using a Web-based Approach.”
- [32] Hokyun Lee, Seungmin Lee, and Taekyong Nam, “Implementation of high performance objectionable video classification system,” in *2006 8th International Conference Advanced Communication Technology*, 2006, p. 4 pp.-pp.962.
- [33] D. Liu, X.-S. Hua, M. Wang, and H. Zhang, “Boost search relevance for tag-based social image retrieval,” in *2009 IEEE International Conference on Multimedia and Expo*, 2009, pp. 1636–1639.
- [34] J. A. Da, S. Júnior, R. E. Marçal, and M. A. Batista, “Image Retrieval: Importance and Applications.”
- [35] S. Badghaiya and A. Bharve, “Image Classification using Tag and Segmentation based Retrieval,” *Int. J. Comput. Appl.*, vol. 103, no. 15, pp. 20–23, Oct. 2014.
- [36] A. N. Bhute and B. B. Meshram, “Text Based Approach For Indexing And Retrieval Of Image And Video: A Review,” Apr. 2014.
- [37] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, 1992, pp. 144–152.
- [38] S. Tong and D. Koller, “Support Vector Machine Active Learning with Applications to Text Classification,” *J. Mach. Learn. Res.*, vol. 2, no. 11, pp. 45–66, 2001.
- [39] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” *R News*, vol. 2, no. 3, pp. 1–10, 2002.
- [40] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [41] Kowsari, Kamran, Mojtaba Heidarysafa, Donald E. Brown, Kiana Jafari Meimandi, and Laura E. Barnes. “RMDL: Random Multimodel Deep Learning for Classification.” In *Proceedings of the 2nd International Conference on Information System and Data Mining*, pp. 19-28. ACM, 2018.
- [42] Kowsari, Kamran, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. “Hdltex: Hierarchical deep learning for text classification.” In *Machine Learning and Applications (ICMLA)*, pp. 364-371, 2017.

Access Control Model for Modern Virtual e-Government Services: Saudi Arabian Case Study

Rand Albrahim, Hessah Alsalamah, Shada Alsalamah, Mehmet Aksoy

Department of Information Systems
King Saud University
Riyadh, Saudi Arabia

Abstract—e-Government services require intensive information exchange and interconnection among governmental agencies to provide specialized online services and allow informed decision-making. This could compromise the integrity, confidentiality, and/or availability of the information being exchanged. Government agencies are accountable and liable for the protection of information they possess and use on a least privilege security principle basis even after dissemination. However, traditional access control models are short of achieving this as they do not allow dynamic access to unknown users to the system, they do not provide security controls at a fine-grained level, and they do not provide persistent control over this information. This paper proposes a novel secure access control model for cross-governmental agencies. The secure model deploys a Role-centric Mandatory Access Control MAC (R-MAC) model, suggests a classification scheme for e-Government information, and enforces its application using XML security technologies. By using the proposed model, privacy could be preserved by having dynamic, persistent, and fine-grained control over their shared information.

Keywords—Access control; cloud infrastructure; data classification scheme; data exchange; e-government; fine-grained access; implementation framework; persistent control; XML security technologies; Saudi Arabia

I. INTRODUCTION

Electronic government (i.e. e-Government) refers to the use of Information and Communication Technologies (ICTs) to provide citizens with access to the country's public services [1]. The aim of e-Government is to improve efficiency [2], reduce the cost of government agencies' processes, and enhance administrative efforts for citizens and businesses; this is done by managing the interacting process with public authorities in a speedy manner [3], and creating a virtual electronic government that leads to economic growth and better transparency [3].

Many developing countries around the globe are shifting towards an electronic service (i.e. e-service) delivery model, and Saudi Arabia is no different. The Saudi government has invested intensively in building the infrastructure for technologies to support e-Government service [4]. However, government agencies do not deliver citizen-centred services for many reasons, and according to studies in [4], [5], information security risks, misplaced trust, privacy issues, and shortages in terms of available infrastructure take precedence. Therefore, there is still a necessity for the government to work harder towards providing tailored e-services. This can be

achieved through a collaboration between its different governmental agencies, both public and private sectors [4], [6].

Government agencies interconnect with each other in different ways, either within the government's premises, across different governmental agencies (G2G), government and business/commerce (G2B), or government and citizens (G2C) [4], [7], [8]. These service delivery models pose risks that can undoubtedly compromise the integrity, confidentiality, or availability of data and information being exchanged [8]. E-Governments have established different ways to communicate in a safe manner. The Saudi Arabian e-Government, like many other developing countries, for example, has established the Government Service Bus (GSB), a G2G that connects all agencies in Saudi Arabia and enables them to exchange information and services in the form of web services [9]. This raises the question of who can see what within the GSB. It is the government's responsibility to protect the personal information they possess and use under law [6]. According to Resolution 40 of the Saudi Ministers' Council, "Information and data relevant to the user or applicant for a government service shall be viewed only by authorized persons" [10]. This rule clearly states that all information systems used for the collection, transformation, processing, and/or manipulation of e-Government information must enforce appropriate information security controls to maintain the right balance between this information's availability, confidentiality, and integrity. This is to ensure the security of those systems' information. Applying the "Least Privilege" access control principle can attain this balance. This principle grants authorised members of the organization access to the absolute minimum amount of information for the absolute minimum amount of time required to complete their duties [11]. To achieve the *least privilege* security principle, an access control mechanism needs to be deployed. However, traditional access control models are, firstly, static and inflexible to grant access to unknown users of the system. Secondly, they are coarse-grained [12] models capable of granting access either to the whole information resource or none of it, since they do not allow access at finer granularity. Finally, they do not provide persistent control over this information [13]. Most of the studies focus on securing access to data instead of securing the data itself [14]. Nevertheless, to preserve ownership over data even when it resides outside the premises, it is important to have continuous protection with information security controls that move along with the data [14] in both the physical and network levels [15]. This also

includes all states of data when stored, processed, transited, and at a final destination [24].

This paper proposes an access control model (named R-MAC) for e-Government's connected web services to the GSB using an access control model that would achieve least privilege principle, and as stated by the Saudi e-government law, it is centred on the security of outsourced governmental data used within GSB to provide specialized online services. R-MAC is a novel approach that incorporates the properties of Mandatory Access Control (MAC) and Role-based Access Control (RBAC) into a new role-centric MAC model and employs XML security technology combined with a data classification scheme suitable for e-government information. MAC is a model in which the security policies and permissions for a subject to access an object are strictly constrained by the system [17], and RBAC grants subjects access to objects based on their role. XML security technologies reuse existing cryptographic and other security technologies whenever possible. It consists of XML digital signature, XML Key Management Specification (XKMS), XML encryption, Security Assertion Mark-up Language (SAML), and XML Access Control Mark-up Language (XACML) [18]. R-MAC provides a secure data exchange framework using some of those components to help preserve the ownership of data at a fine granularity. Fine-grained access control provides the right privileges to a user to grant him/her access to an asset only if this user is authorized [19]. In addition, XML security technologies provide fine-grained and persistent security controls that move with the data. Achieving a safe platform for data exchange in e-Government services enables dynamic, persistent, and fine-grained control for specialised online services through the collaboration of different government agencies.

The remainder of this paper is organized as follows. Section II provides a background of the Saudi e-government program and the use of web services. After that Section III presents the literature review. Then, we present our methodology in Section IV. Finally, the conclusion is provided in Section V.

II. BACKGROUND

Saudi Arabia's Vision 2030 [20] is a plan adopted by the Saudi Government to guide economical and developmental action in Saudi Arabia [20]. The *National Transformation Program 2020* [21] was launched across 24 governmental bodies across Saudi Arabia to help achieve the ambitious goals of *Saudi Arabia's Vision 2030*. The Ministry of Communications and Information Technology [22] set a number of strategic objectives that correspond to relevant vision 2030 objectives, which include: "*Strategic Objective 3: Develop and activate smart government transactions based on a common infrastructure*" [21].

According to the United Nations' index for the development of e-Government, Saudi Arabia is currently ranked 36 globally and the target rank by 2020 is 25 [23]. The current maturity level of the government services transformation to e-services is 44% and the aim is to reach 85% by 2020 [21]. However, there are a number of reasons

that hinder the adoption of e-Government services in general regardless of its model. These include, but are not limited to, poor skills, technology literacy [12], and security and privacy concerns regarding their shared personal information [24] (i.e. personal, financial, and medical data [25]). Moreover, the challenges related to the adoption and implementation of the Saudi e-Government are specific to infrastructure cost, computer literacy, accessibility, availability, trust, and privacy issues [4].

YESSER [34] is one of the key national programs in Saudi Arabia that specifically enables the delivery of e-Government services across government agencies. This is achieved through the development of a number of interrelated initiatives, for interoperability and networking. First and foremost, YESSER deploys an interoperability framework (YEFI) [26] that defines the set of policies to be implemented by government agencies to ensure a standardization of information and service exchange. It also defines the data types, schema, meta-data elements, dictionaries, and technical policies. The technical policies include the integration approach and set of standards, for connectivity, security, and information access and delivery [26]. Second is the GSB [9] which is the central platform of integration and services for government e-services and transactions. The GSB provides support for web services [9] as illustrated in Fig. 1. Currently, there are 69 agencies connected to the GSB providing 115 different services [9].

A Web service is standardized method, which allows different systems to communicate over a network. It can be a user requesting a service from a Web server or a Web server requesting a service from another Web server [27]. Web services have many advantages, including multiple heterogeneous platform compatibility [28], language-independency [36], increased information availability and ease of access [29], and maintaining up-to-date data. However, Web services do not have any predefined security model, and therefore require the additional implementation of techniques to protect exchanged information [28], as well as the deployment of a framework that enforces a strong security architecture [30].



Fig. 1. GSB Overview.

Web services interact using Simple Object Access Protocol (SOAP) messages [27]. SOAP is a standard for one-way and request-response messaging transmitted over HTTP protocol. To protect SOAP messages, Advancing Open Standards for the Information Society (OASIS) set web service standards called Web Service Security (WS-Security), which is a message level security mechanism that consists of digital signature and encryption techniques [27]. SOAP messages are based on Extensible Mark-up Language (XML) data format. XML, on the other hand, is used in many areas to store, retrieve, and provide data and information in an organized format [31] and it is considered one of the most extensively used data exchange languages across the internet [32] because of its immense compatibility in transit [33]. There are many advantages in using XML data representation, such as: the notions of elements, it is extensible, it allows the separation of display and content and it can present complex structure in an easy way [32]. Nevertheless, XML falls short of guaranteeing the security by itself, and hence, a secure application framework is needed as a precondition to have it programmed as needed [32].

III. LITERATURE REVIEW

A. Related Work on Security of E-Governments

Technologies that have been used to maintain security in e-Governments include one-time passwords, cryptography, firewalls, analysis tools, and monitoring tools [34]. The report in [35] introduced a data classification for the e-Government of the United Kingdom which includes three security classifications (OFFICIAL, SECRET, and TOP SECRET) to indicate the level of data sensitivity and to specify how to handle personnel security, physical security, and information security over each data classification type.

The authors in [36] introduce a security model for the e-Government in United Arab Emirates (UAE) that is based on Public Key Infrastructure (PKI) (certificate and digital signature), biometrics (finger print), and hardware security tools (Tokens). Another security model for e-Government was introduced by [37] which is also based on PKI in addition to SSL channel. A proposed design for a framework for the Sudanese e-Government was introduced by [38] which suggests that the technical layer should include: an access control mechanism, authentication and password, cryptography, the use of tamper resistance protocol, a secure communication link, analysis tools, monitoring tools, bandwidth techniques, validate and filter input, and a one-time password. Another study was conducted by [39] which introduced an information security governance model for e-services in South African developing countries e-Government projects which suggests that the operational layer should include an identity management framework for authentication and authorization and a new token-based technique for implementing identity management. Their study also used DES for encryption/decryption process.

B. Related Work on XML Security Technologies

The notion of Web services has been absolutely crucial in the IT industry. Currently, all business transactions depend on Web services to achieve their desired goals [40].

However, the security of Web services is an emerging topic of discussion. To secure Web services, it is essential to secure their content which is based on XML language. XML security technologies reuse existing cryptographic and other security technologies whenever possible. It consists of XML digital signature, XKMS, XML encryption, SAML and XACML. Some of the methods used to secure XML-based Web services were introduced by [15] and the authors present the use of XML Signature to ensure the integrity and XML encryption to provide confidentiality for XML messages in Web services. In [41], the authors evaluated authentication, authorization, integration, confidentiality, and non-repudiation when using XML encryption and XML Signature in web-services in an e-business scenario and proved that all those parameters could be guaranteed when combining both XML encryption and XML signature. Meanwhile, in [33], the authors examined how XML Web security could provide privacy, certification, and integrity. They applied XML encryption and XML signature for data and messages in transaction and in storage. It was found that not only were security requirements established, but also the performance of Search Engine Optimization (SEO) was enhanced with the parsing of descriptive tags rather than unstructured data.

XML encryption and XML signature are both low level features to make the data itself secure [16]. On the other hand, Access Control is considered a high-level approach to security policies that provide secure access to data and both encryption and signature are designed to handle communication security [31]. The authors in [32] proposed an XML access control to guarantee application safety and they proved that the advantages of using XML access control are a fine-grained access on an element level, the use of different safety strategies over different parts on the same document, the use of a safety process for encryption and digital signature, and protecting network resources.

C. Related Work on Access Control

Access control is one of the key aspects of information security [42]. Access Control is a mechanism to provide privileges to a user to access a particular asset, only if this user is authorized [19] and Access Control is domain specific. Methods that are widely used in Access Control mentioned in [17], [43] are: DAC, MAC, and RBAC. There are many studies in the literature on access control methods. A study related to DAC was presented in [44] regarding its complexity, safety, and issues in object-oriented databases. RBAC was widely studied as well, and researchers in [45] proposed object sensitive role assignment, which is a generalized RBAC model for object-oriented languages.

However, much research has focused on exploring ways in which they can integrate different Access Control models to achieve better security and efficiency [17]. In [46], the authors introduced a model which combines MAC and ABAC, retaining the strict nature of MAC approach and providing more access control decisions in attributes. In [47], the authors introduced a model that consists of two layers: one layer is called the “aboveground” level and it is a traditional RBAC that is extended with environment constraints, while a second layer, called the “underground” layer, focuses on constructing attribute-based policies to automatically create primary RBAC

model on the “aboveground level”. This model combines both aspects from RBAC and ABAC. Authors in [48] also proposed a model that combines features from ABAC and RBAC to provide effective access for application where there are static attributes such as qualification and city, and dynamic attributes such as the time of the day. Authors in [17] mentioned that if attributes in ABAC were chosen appropriately, they can capture the identities and access lists (DAC), security labels, clearance and classifications (MAC) and roles (RBAC). Hence, ABAC supplements those current dominant models rather than substituting them. The authors in [49] introduced RABAC, which is a Role-centric Attribute-based Access Control and this is an extension to NIST RBAC with policies for permission filtering to overcome the issue of role explosion. Other studies that were introduced to parameterize RBAC were introduced by [50] and [51]. A study by [52] proposed an attribute-based constraint specification language to express constraints in a way that it can be assigned to the attributes of different entities.

Access control for web services has been an issue that was studied in the literature as well. The authors in [53] introduced an approach to handle authorization of web requests in web services based on the concept of identity tracking and access percentile of the invoking of the web service. The authors in [54] argued that the two main issues that need to be addressed in the access control of web services are, firstly, restricting the access to authorized people, and, secondly, protecting the integrity and confidentiality of XML messages exchanged through web services. However, relying on security techniques currently used in web services such as HTTPS (HTTP over Secure Sockets Layer Protocol) cannot provide for example authorization to regulate the actions of users by allowing or disallowing an operation. Researchers in [55] proposed Authorization-based Access Control (ABAC) URL that is compatible with common web tools. A web service access control scheme was proposed in [56] where the access control scheme incorporates user password and web server log, and it grants access based on the user access behaviour by tracking the web access history. The access is granted based on the user password, date of last request, page visited (URL), and status action, association rules mining and PrefixSpan algorithms are used to match the active users’ access pattern with the user access data discovered from the user access history before being analysed to make the access decision. The authors in [57] proposed an Access Control model for information retrieval in EHR (Electronic Health Record) systems where the patient is allowed to define the access rules concerning their clinical information. The aim of their model is to increase the confidentiality and integrity of the data and raise the patients’ trust in the EHR systems. A study by [58] was conducted on Privacy-aware access control model and their use in web services. Although the generalization of data can guarantee user privacy, the over generalization of data may result in useless data, so to guarantee the right balance between data usability and the disclosure of privacy, the authors analyse how to manage an effective access process through a trust-based decision and ongoing access control policies. The authors in [59] proposed a generic access control model for the cloud that can be used with different cloud

service models and it is based on Kerberos as well as access control lists and authorization tickets.

The overview of related work presented shows that many recent studies on access control focused on the field of Web services in many domains. Finding the best access control model for specific and generic domains is an emerging and current topic. Therefore, the research in this paper analyses the best access control framework that is suitable within the domain of e-Government and examines this in a real case scenario to prove its feasibility.

IV. METHODOLOGY

This research follows the steps of Design Science Research Methodology (DSRM) to present the research design. DSRM is one of the mainly used forms of methodologies in the field of Information Systems (IS); it includes the construction of new knowledge through the design of novel or innovative artefacts [60]. The Design Science process includes six steps which are outlined next.

A. Problem Identification and Motivation

The problem includes overcoming the limitations of traditional access control models and help achieve the principle of least privilege security principle.

B. Definition of the Objectives for a Solution

To define the objectives for a solution, we have explored the challenges that face the successful adoption of the e-Government program in Saudi Arabia. Researchers in [4], [61], [62] have identified security and privacy as one of the main obstacles. This is in addition to other obstacles, such as the establishment of the infrastructure, availability, computer literacy, trust, accessibility, authentication, usability, and accountability. Having security and privacy as a main issue has led (YESSER) to limit the data and services shared to preserve the confidentiality and to avoid privacy violations.

In order to achieve the above goal for e-Government Web services, specific objectives of this paper are set as follows:

- Apply a data classification scheme for the information being exchanged within and outside the GSB.
- Develop a security intermediary between the service requester and the service provider to provide authentication and authorization.
- Utilize XML security technology to provide security controls that are both fine-grained and persistent.
- Implement a case study of a real scenario within GSB using the proposed security model.

C. Design and Development

Our proposed solution is an Access Control model that will serve as a security intermediary that will intercept any access request to a web service and convert it to an authorization request to determine the suitable response output. R-MAC is a role-centric MAC model that incorporates the properties of MAC and RBAC for granting access permissions by giving clearances to roles rather than individual users to provide more flexibility and better

expression of application-level security. It also utilizes XML security technologies to achieve persistent and fine-grained security over the information, even when it is outside the physical premises.

To implement the proposed model, first, we apply a dynamic role assignment. For our case study, which is the e-Government of Saudi Arabia, the role assignment within the GSB is determined by YESSER. In addition, YESSER provides Authorization using one-time password [63]. After this, we deploy a MAC model by assigning the following:

1) *Classification for objects*: We introduce a data classification scheme for information that is based on the Saudi e-Government security law. This classification scheme adopts the Traffic Light Protocol since it is one of the well-known data classification schemes and is widely used in different domains and systems. Traffic Light Protocol employs four colours to indicate the level of data sensitivity and the sharing boundary to be applied on recipients. For example, if the data is classified as Red, only the users that are given the clearance Red will be able to view and modify. For our case study, which is the e-Government of Saudi Arabia, the information classification is based on the Saudi e-Government law. Table I presents a summary of the Saudi e-Government law [64] that specifies the corresponding security control.

2) *Clearances for subjects*: An attribute is added to the roles to specify the clearance level and this is dynamically performed after the authentication step. After that, the access control model will perform a role-to-permission assignment.

By applying the Read-Down rule used in MAC [31], a subject with Red clearance can view all data classified as Red, Amber, Green, and White. A subject with Amber clearance can view all data classified as Amber, Green, and White. A subject with Green clearance can view all data classified as Green and White. A subject with White clearance can view all data classified as White. Fig. 2 illustrates this rule:

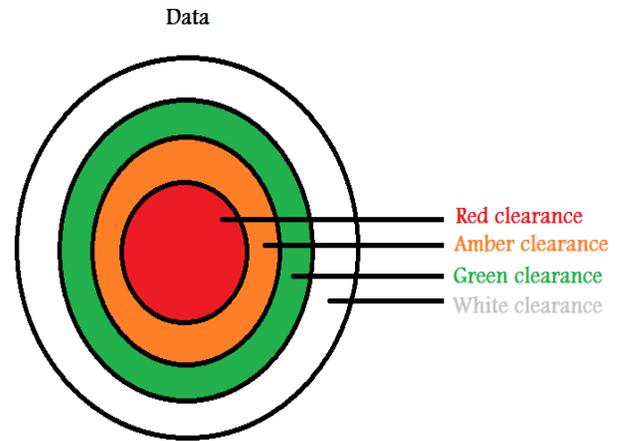


Fig. 2. Read-Down Rule.

Applying an access control mechanism that combines features from MAC and RBAC in addition to securing the data itself with XML encryption illustrated in the architecture in Fig. 3 will help overcome the limitations of traditional access control models and provide a safe platform for data exchange and distribution using web services.

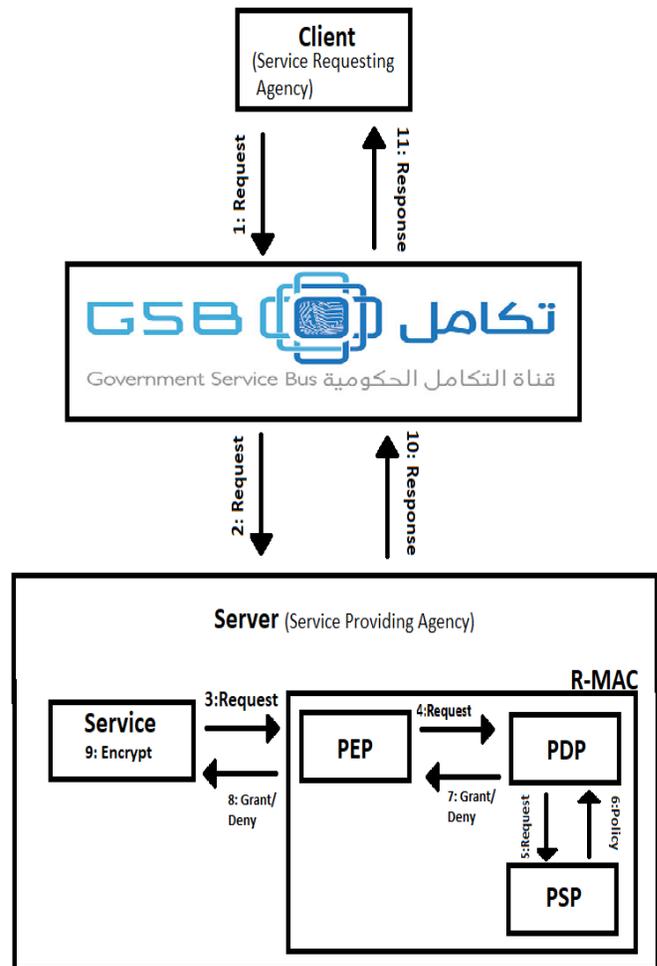


Fig. 3. R-MAC Architecture.

TABLE I. DATA CLASSIFICATION IN E-GOVERNMENT

Colour	Security Control	Description
Red	Top confidential	Information that may cause damage to the security of the state. Such information may only be accessed by senior officials.
Amber	Very Confidential	Information that may cause damage to public or private interests. This information is only available to specialists.
Green	Confidential	Information that relates to individual cases and may have a negative impact on the social life of the community or individuals.
White	Disclosure is not limited	Non-confidential public information.

R-MAC model works as an intermediary that will intercept a business request and convert it into an authorization request to provide authentication and authorization for the web service through the following steps:

- 1) Starts when a business request is made to a service between the agencies connected to the GSB.
- 2) GSB managed by YESSER will provide Authentication using one-time password and role assignment for the service requesting agency and connect it with the service providing agency which acts as the server.
- 3) The business request is intercepted by an intermediary which is the proposed access control model (R-MAC) and converted to an authorization request by the Policy Enforcement Point (PEP).
- 4) The authorization request is then sent to the Policy Decision Point (PDP) to evaluate it and return the decision to the PEP.
- 5) To evaluate the authorization request, PDP examines the policies saved in the Policy Storage Point (PSP).
- 6) The suitable policy is then sent back to the PDP to Grant/Deny the access.
- 7) The decision is then sent to the PEP to enforce it over the service response.
- 8) The suitable response is then formatted.
- 9) The response is encrypted using XML encryption to guarantee persistent security control through the transmission of information.
- 10) The suitable response is then returned to the GSB.
- 11) The client receives the response and decrypts it using the appropriate key.

D. Demonstration

To illustrate the proposed access control model, a scenario is provided where RMAC can be implemented to help preserve the confidentiality of data and help achieve the principle of Least Privilege specified by the Saudi law. TAKAFUL charity organization was carefully studied as an exemplar.

TABLE II. INFORMATION REGARDING THE WEB SERVICE

Service Name	Citizen Profile
Operation Name	Retrieve the personal profile of the citizen
Service Provider	Ministry of Interior
Service Type	Information Retrieval
Process Description	The process retrieves the full profile of the entered citizen
Input	<ul style="list-style-type: none">• ID Number
Output	<ul style="list-style-type: none">• Full name (name)• ID information (ID)• Location of Birth (LOB)• Birth Certificate Number (BCN)• Gender• Social status (Sstatus)• Job status (Jstatus)• Life and death status (Lstatus)

TAKAFUL [65] is a charitable organization which helps under-privileged students by providing financial and psychological support. TAKAFUL organization connects with other governmental agencies and data sources to gather information regarding applicants to determine their eligibility such as the Ministry of Interior, the Ministry of Civil Services, General Organization of Social Insurance, Public Pension Agency, Ministry of Commerce and Investment, Ministry of Labor and Social Development, Ministry of Justice and the Ministry of Education.

Our chosen scenario, which is a part of the Eligibility Process, is the interaction between TAKAFUL as a data consumer and the Ministry of Interior (MOI) as a data owner in order to check the status of the parents. Table II provides information regarding this service.

The previous Web service from Table II was implemented where the proposed access control model serves as a security intermediary between the service providing agency and other agencies connected to the GSB. It will deploy a MAC model where the access rights are constrained by the system based on a data classification scheme. The policies were specified using XACML as a separate component to give it more flexibility since data and users can be updated without affecting the policies. Finally, based on the clearance level, classified information is displayed.

E. Evaluation

The evaluation and validation of the proposed model have two dimensions, which are as follows:

1) Testing the model with test cases and comparing the expected output with the actual output to determine that the model works as expected. The test cases proved that the expected output matches the actual output.

2) Checking the value and usability aspects of the model by distributing the validation form developed in [66] to help the participants in selecting the rate of validity from different success standards. The validation form was distributed among five highly recognized security practitioners in e-Government. The results of the analysis confirmed the standards contained in the success of the proposed model. However, some of the suggestions include utilizing a robust authentication mechanism that is not weaker than 2-factor authentication: for instance, a strong password and hardware token.

The advantages of adopting the proposed model include:

- Applying a Role-Centric MAC model provides more flexibility to it since the clearances will be given to roles rather than individual users.
- Adopting our proposed data classification scheme and role clearance provides a fine-grained control to enforce the principle of Least Privilege.
- Utilizing XML security technologies provides a persistent and fine-grained end-to-end security control even when the information is outside the physical premises.

F. Communication

The final step of DSRM is publications in academic journals and professional outlets.

V. CONCLUSION

The security of e-Government services is one of the major concerns nowadays, especially in terms of the confidentiality of data owned by the e-Government agency. In order to provide specialized online services, governments must interconnect and exchange pieces of information to paint the full picture and make informed decisions. This exchange of information can compromise the integrity, confidentiality, or availability of that information. This paper proposes an access control model that overcomes the limitations of traditional access control models by combining features from MAC and RBAC and by giving clearances to roles rather than individual users to give it more flexibility and better expression of application-level security. This paper also introduces a data classification scheme that will help preserve the security of the information being exchanged within and outside the GSB by providing a fine-grained access control model that complies with the Saudi law which strictly grants access on a Least Privilege security principle basis that enables fine-grained access control. In addition, XML security technologies are utilized to achieve persistent and fine-grained control over the data even when it resides outside the physical premises. The proposed access control model, which uses the combined R-MAC model along with the suggested classification scheme and enforces it through the XML security technology, is novel. The information classification scheme with the corresponding clearance levels proposed in this work is the result of an analysis of the Saudi e-Government law. The proposed model was evaluated with a case study of the interaction conducted through the GSB and it shows that the principle of Least Privilege is enforced, and the security of data is preserved.

The presented work in this paper provides the basis for accomplishing a secure access control model that can provide flexible, fine-grained, and persistent access control for the information shared in Saudi e-Government interaction. However, the novel access control model proposed could be applied in any collaborative cloud-based environment with its own data classification scheme to address the limitations in the current security models. Moreover, other classification schemes and role clearances could be investigated to achieve different granularity of control.

Another aspect that is worthy of further study relates to the role assignment process. Currently, the role assignment process in the e-Government of Saudi Arabia is performed by the administration of YESSER. However, in the future, it would be beneficial to introduce a specific mechanism for the role assignment process since accurate role assignment is a key to preventing privacy violations.

The proposed utilization of XML security technologies presented in this research includes a basic symmetric key encryption for the XML formatted documents. However, in the future, it is crucial to establish a Public Key Infrastructure (PKI) to manage digital certificate and public-key encryptions.

ACKNOWLEDGMENTS

This research was sponsored by the Deanship of Graduate Studies at King Saud University.

REFERENCES

- [1] F. Sá, Á. Rocha, and M. P. Cota, "Potential dimensions for a local e-Government services quality model," *Telemat. Informatics*, vol. 33, pp. 270–276, 2016.
- [2] A. M. Al-Khoury, M. Farmer, and J. Qadri, "A Government Framework to Address Identity, Trust and Security in E-government: the Case of UAE Identity Management Infrastructure," *Eur. Sci. J.*, vol. 10, no. 10, pp. 85–98, 2014.
- [3] H. Ritchi, I. Wahyudi, and A. Susanto, "Research Program on Key Success Factors of e-Government and Their Impact on Accounting Information Quality," in *2nd Global Conference on Business and Social Science*, 2015, vol. 211, pp. 673–680.
- [4] S. S. Basamh and H. A. Qudaih, "E-Government Implementation in the Kingdom of Saudi Arabia: An Exploratory Study on Current Practices, Obstacles & Challenges," *Int. J. Humanit. Soc. Sci.*, vol. 4, no. 2, pp. 296–300, 2014.
- [5] E. Nyakwende and A. Al Mazari, "Factors Affecting the Development of e-Government in Saudi Arabia," in *International Conference on Electronic Government and the Information Systems Perspective*, 2012, pp. 19–28.
- [6] P.-L. Sun, C.-Y. Ku, and D.-H. Shih, "An implementation framework for E-Government 2.0," *Telemat. Informatics*, vol. 32, no. 3, pp. 504–520, 2015.
- [7] M. Zubi and H. Alonaizat, "E-government and Security Requirements for Information Systems and Privacy(Performance Leakage)," *J. Manag. Res.*, vol. 4, no. 4, pp. 367–375, 2012.
- [8] R. G. Hassan and O. O. Khalifa, "E-Government- An Information Security Perspective," *Int. J. Comput. Trends Technol.*, vol. 36, no. 1, pp. 1–9, 2016.
- [9] National Enterprise Architecture Office Management at Yesser, "National Application Reference Model," e-Government Program (Yesser), 2015.
- [10] YESSER, "YEFI - Data Standards Catalogue," Kingdom of Saudi Arabia, 2008.
- [11] M. E. Whitman and H. J. Mattord, *Management of information security*. Stamford: Cengage Learning, 2014.
- [12] S. Saha, D. Bhattacharyya, T. Kim, and S. K. Bandyopadhyay, "Model Based Threat and Vulnerability Analysis of E-Governance Systems," *Int. J. u- e- Serv. Sci. Technol.*, vol. 3, no. 2, pp. 7–22, 2010.
- [13] S. Al-Salamah and M. J. Hilton, "Towards Information Sharing in Virtual Organisations: The Development of an Icon-based Information Control Model," Cardiff University, United Kingdom, 2009.
- [14] S. Alsalamah, A. Gray, J. Hilton, and H. Alsalamah, "Information Security Requirements in Patient-Centred Healthcare Supporting Systems," in *14th World Congress on Medical and Health Informatics (Medinfo)*, 2013, pp. 812–816.
- [15] G. Yue-sheng, Y. Meng-tao, and G. Yong, "Web Services Security Based on XML Signature and XML Encryption," *J. Networks*, vol. 5, no. 9, pp. 1092–1097, 2010.
- [16] M. Zubi and H. Alonaizat, "e-Government and Security Requirements for Information Systems and Privacy(Performance Leakage)," *J. Manag. Res.*, vol. 4, no. 4, pp. 367–375, 2012.
- [17] Xin JinRam, KrishnanRavi, and Sandhu, "A Unified Attribute-Based Access Control Model Covering DAC, MAC and RBAC," in *IFIP Annual Conference on Data and Applications Security and Privacy*, 2012, vol. 7371, pp. 41–55.
- [18] B. P. Verma, S. Kumar, and P. Sharma, "A novel approach for Multi-Tier security for XML based documents," *IOSP J. Comput. Eng.*, vol. 5, no. 4, pp. 1–4, 2012.
- [19] S. Chatterjeea and T. Sarmah, "An Efficient Fine Grained Access Control Scheme Based On Attributes For Enterprise Class Applications," *2014 International Conference on Signal Propagation and Computer Technology (ICSPCT)*. pp. 273–287.

- [20] Saudi Government, "Saudi Arabia's Vision 2030," 2016. [Online]. Available: <http://vision2030.gov.sa/en/ntp>.
- [21] YESSER, "National transformation program 2020," 2016. [Online]. Available: http://vision2030.gov.sa/sites/default/files/NTP_En.pdf.
- [22] Ministry of Communication and Information Technology, "Ministry of Communication and Information Technology," MCIT, 2017. [Online]. Available: <http://www.mcit.gov.sa/En/Pages/default.aspx>.
- [23] Ministry of Interior, "Ministry of Interior," 2017. [Online]. Available: <https://www.moi.gov.sa/wps/portal/Home/Home>.
- [24] G. Tokdemir and Y. Paçin, "Adoption of e-government services in Turkey," Elsevier Comput. Hum. Behav., vol. 66, pp. 168–178, 2016.
- [25] K. K. Smitha and K. Chitharanjan, "Security of Data in Cloud based E-Governance System," Spec. Issue Int. J. Comput. Appl. Adv. Comput. Commun. Technol. HPC Appl., pp. 1–6, 2012.
- [26] YESSER, "YEFI – Yesser Framework For Interoperability," 2005.
- [27] D. Jamil and H. Zaki, "Security Issues in Cloud Computing and Countermeasures," Int. J. Eng. Sci. Technol., vol. 3, no. 4, pp. 2672–2676, 2011.
- [28] N. Dilber, "Restful web services security by using ASP.NET web API MVC based," J. Indep. Stud. Res., vol. 12, no. 1, pp. 4–10, 2015.
- [29] K. S. Tharun, M. Prudhvi, and S. S. Reddy, "Advantages of WCF Over web services," Int. J. Comput. Sci. Mob. Comput., vol. 2, no. 4, pp. 340–345, 2013.
- [30] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," J. Netw. Comput. Appl., vol. 34, pp. 1–11, 2011.
- [31] N. S. Farooqi and D. S. North, "Applying Dynamic Trust Based Access Control to Improve XML Databases Security.," University of Sheffield, 2013.
- [32] H. Zhang, Q. Guan, and W. Luo, "The Study of Access Control Model Using XML," Int. J. Secur. Its Appl., vol. 9, no. 7, pp. 179–188, 2015.
- [33] R. Menaka and B. Ashadevi, "Survey on Signed Xml Encryption for Multi-Tier Web Services Security," Indian J. Sci. Technol., vol. 9, pp. 1–10, 2016.
- [34] S. Singh and S. Karaulia, "E-Governance: Information Security Issues," in International Conference on Computer Science and Information Technology (ICCSIT'2011), 2011, pp. 120–124.
- [35] cabinet office, "Government Security Classifications," 2014.
- [36] J. T. Jaafar, N. Hamza, and N. Hamza, "Security Model in E-government with Biometric based on PKI," Int. J. Comput. Appl., vol. 93, no. 6, pp. 33–39, 2014.
- [37] W. Zhong, "Research On E-Government Security Model," in International Conference on E-Business and E-Government, 2010, pp. 699–702.
- [38] O. Ali, "A proposed Design of a Framework for Sudanese E-Government Security Model," Sudan University of Science and Technology, 2017.
- [39] A. Ramtohol and S. Soyjaudah, K. M., "Information security governance for e-services in southern African developing countries e-Government projects," J. Sci. Technol. Policy Manag., vol. 7, no. 1, pp. 26–42, 2016.
- [40] S. Gadwar and D. Sable, "Securing Web Services Based on XML Signature and XML Encryption," Int. J. Res. Advent Technol., vol. 2, no. 2, pp. 1–5, 2014.
- [41] G. Abraham, Krishnakumar, Venkatasubramanian, and K. Borasia, "Securing Web Services Using XML Signature and XML Encryption," School of Computer Science and Engineering, VIT University, Vellore, India, 2013.
- [42] S. Gostojić, G. Sladic, B. Milosavljević, and Z. Konjovic, "Context-Sensitive Access Control Model for Government Services," J. Organ. Comput. Electron. Commer., vol. 22, no. 2, pp. 184–213, 2012.
- [43] S. Nagaraju, L. Parthiban, and S. Kumar, "An Enhanced Symmetric Role-Based Access Control Using Fingerprint Biometrics for Cloud Governance," Parallel Cloud Comput. Res., vol. 1, no. 2, pp. 11–16, 2013.
- [44] S. Dranger, R. Sloan, and J. Solworth, "The complexity of discretionary access control," Proceeding IWSEC'06 Proc. 1st Int. Conf. Secur., pp. 405–420, 2006.
- [45] J. Fisher, D. Marino, R. Majumdar, and T. Millstein, "Fine-Grained Access Control with Object-Sensitive Roles," in European Conference on Object-Oriented Programming, 2009, pp. 173–194.
- [46] L. Kerr and J. Alvis-foss, "Combining Mandatory and Attribute-Based Access Control," in 49th Hawaii International Conference on System Sciences (HICSS), 2016, pp. 2616–2623.
- [47] Y.-H. Chen, C.-H. Lu, and P.-Y. Hsu, "Multilayered Information Encryption Scheme with Fine-grained Authentication," in Proceedings of APSIPA Annual Summit and Conference 2015, 2015, pp. 1126–1130.
- [48] D. R. Kuhn, E. J. Coyne, and T. R. Weil, "Adding Attributes to Role-Based Access Control," IEEE Comput., vol. 43, no. 6, pp. 79–81, 2010.
- [49] X. Jin, R. Sandhu, and R. Krishnan, "RABAC: role-centric attribute-based access control," in Proceedings of the 6th international conference on Mathematical Methods, Models and Architectures for Computer Network Security: computer network security, 2012, pp. 84–96.
- [50] A. Abdalah and E. Khayat, "A Formal Model for Parameterized Role-Based Access Control," in Formal Aspects in Security and Trust, 2004, pp. 233–246.
- [51] M. Ge and S. Osborn, "A Design for Parameterized Roles," in Research Directions in Data and Applications Security XVIII, 2004, pp. 251–264.
- [52] K. Bijon, R. Krishnan, and R. Sandhu, "Towards An Attribute Based Constraints Specification Language," in international conference on Social Computing (SocialCom), 2013, pp. 108–113.
- [53] R. Nath and G. Ahuja, "An Authorization Mechanism for Access Control of Resources in the Web Services Paradigm," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 6, pp. 36–42, 2011.
- [54] C. A. Ardagna, "A Web Service Architecture for Enforcing Access Control Policies," Electron. Notes Theor. Comput. Sci., vol. 142, no. 3, pp. 47–62, 2006.
- [55] G. Swamynathan, T. Close, and S. Banerjee, "Scalable Access Control For Web Services," in the fifth international conference on Creating, Connecting and Collaborating through Computing, 2007.
- [56] S. Elsheikh, "Access control scheme for Web services (ACSWS)," in international conference on Computer and Communication Engineering, 2008.
- [57] M. Sicuranza, A. Esposito, and M. Ciampi, "An access control model to minimize the data exchange in the information retrieval," J. Ambient Intell. Humaniz. Comput., vol. 6, no. 6, pp. 741–752, 2015.
- [58] M. Li, X. Sun, H. Wang, Y. Zhang, and Z. Ji, "Privacy-aware access control with trust management in web service," World Wide Web, vol. 14, no. 4, pp. 407–430, 2011.
- [59] H. Kaffel-Ben Ayed and B. Zaghdoudi, "A generic Kerberos-based access control system for the cloud," Ann. Telecommun., vol. 71, no. 9–10, pp. 555–567, 2016.
- [60] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," J. Manag. Inf. Syst., vol. 24, no. 3, pp. 45–78, 2007.
- [61] J. Al-Khouri, A. M., & Bal, "Electronic Government in the GCC Countries," Int. J. Soc. Sci., vol. 1, no. 2, pp. 83–98, 2007.
- [62] O. Alshehri, M., Drew, S., & Alfarraj, "A Comprehensive Analysis of E-government services adoption in Saudi Arabia: Obstacles and Challenges.," High. Educ., vol. 8, no. 2, pp. 1–6, 2012.
- [63] e-G. P. (YESSER), "No Title," Single sign-on (SSO), 2017. [Online]. Available: https://www.yesser.gov.sa/EN/buildingblocks/pages/the_single_sign-on.aspx.
- [64] National Center for Documentation and Archives, "List of Documents to be Consulted and Circulated," 2004.
- [65] TAKAFUL Charity Foundation, "TAKAFUL," 2017. [Online]. Available: <https://www.takaful.org.sa/>.
- [66] M. Lankhorst, Enterprise Modelling, Communication and Analysis, vol. 2. London, United Kingdom, 2009.

Evaluation of the Impact of Usability in Arabic University Websites: Comparison between Saudi Arabia and the UK

Mohamed Benaida, Abdallah Namoun, Ahmad Taleb

Faculty of Computer and Information Systems
Islamic University of Madinah, Saudi Arabia

Abstract—Today usability is a crucial factor that can affect any website. The purpose of this study is to explore major usability defects within Saudi university websites in comparison to British university websites from a Saudi student perspective. In addition, students are expected to achieve their goal when surfing a Saudi Arabian university website comfortably and efficiently without any complication. This study uses two methods to evaluate and measure usability problems; user testing and thinking aloud. Both methods are very useful and effective for collecting data from participants. Based on the ranking of the universities, 60 students were split evenly into three groups; each group was asked to evaluate a different pair of university websites from different ranking levels, one from the UK and the other from KSA. The evaluation performed by each group was gathered using the SUS (System Usability Scaling) questionnaire to find flaws within the usability of the website. During the experiment, the participants' opinions were collected using the thinking aloud method. The findings of this research showed that all Saudi universities in all tiers had significant problems within the usability of their websites. The most frequent problems found were, inconsistency, integration, confidence and satisfaction. Other less frequent problems that were found during this study were design concepts, easy use of websites and comfort of students. Saudi universities can learn from the differences in the quality between both sides to upgrade and redesign their website to achieve user satisfaction, therefore increasing the confidence of the users.

Keywords—Usability; usability evaluation; factor analysis; student satisfaction

I. INTRODUCTION

Students around the world today are more involved in using the internet than ever before. Nowadays, the internet has become the main core of education and the most significant characteristic that can influence the level of knowledge of any educational sector [37]. In the last two decades usability has become a crucial factor that has affected the quality and satisfaction of the users of websites [21], [23], [43]. Web design, information and system quality are variables that can deeply affect the success of usability within a website from the users' perspective [1], [34], [35], [45]. This study has allowed us to enhance the gratification of usability of educational websites in Saudi Arabia.

The contribution of this study can be summarised in several points; firstly, the lack of studies in the field of Arabic usability within educational websites is due to studies being in

the initial stages, therefore this study has come to fill this gap. Besides that the comparison conducted in this study between developed and developing countries provide a clear picture about the level of educational websites in Saudi Arabia and also discover the main barriers that can influence the satisfaction of the users of Arab educational websites. Finally, this study suggests adequate solutions that are revealed by the end users which make this study more reliable.

The layout of this study is as follows: an introduction to the study was given in Section I. Deep literature review is conducted in Section II to explore previous education websites usability. The process of data collection was highlighted and a constructive method was applied in Section III. Results are presented in Section IV, evaluated in Section V and a conclusion is formed in Section VI. The final two sections are "Limitations of Study" and "Future Recommendations".

II. LITERATURE REVIEW

Usability has different definitions based on the field of study. The standard definition of usability according to the international standard organisation [5] is "effectiveness, efficiency and satisfaction with which a specified set of users can achieve, in a specified set of tasks in a particular environment". The author in [10] divides usability into five main factors:

- Learnability: the users should find the system easy to use and complete their task quickly
- Efficiency: the number of tasks that the user can successfully accomplish by using the system
- Memorability: the user can easily remember the system
- Errors: the user can easily recover from a system error
- Satisfaction: the user should feel pleasant when using the system.

A few studies have analysed Arabic educational usability websites. The consistency in Saudi Arabian multilingual websites was examined, one website (King Faisal University) was tested as a case study, several problems were identified, and solutions have been suggested [29]. Other studies focus on the level of usability in Jordanian university websites, the main problems faced on Jordanian websites were related to most areas in usability, such as the design of websites [19],

[20]. Moreover, some studies [18] found that the participants failed to complete any tasks on the websites examined due to the 26 usability problems discovered which is appalling. In this study, the author explores usability problems in the Arab North African educational websites, the satisfaction of Arabic users is still so far. The websites failed to satisfy all usability variables including slow loading speed and high number of HTML objects [24]. In addition, the level of usability in a non-Arabic country was investigated [3], [40] in Turkish universities, the users found the university websites easy to use and that the websites proved to be a very useful source of information regarding the university. Arabic universities suffer from poor usability, effectiveness and learnability which are all very important to provide a smooth and spontaneous experience of these e-learning webpages [25].

Technology and computing plays a major role on the learning of students, especially in universities. Factors like presentation, collaboration and creativity are heavily influenced by technology and usability in most universities, but the use of technology should be made a mandatory part of other teaching and learning processes as well [26]. The way the users perceive websites in general and e-commerce websites in particular, are greatly affected by usability and the culture they belong to. The impact of usability on e-learning systems was analysed by [30] in which they studied a specific e-learning system and with the help of various parameters (for example participation of students in forums, blogs and messaging systems), they evaluated the performance of the system and how it impacts the students. The usability of their selected system was measured using various tests like Technology Acceptance Model and they included students from different universities. They concluded that forums had more impact on the learning of students. Recent studies regarding educational websites concentrate on the localisation and globalisation of multilingual websites, [22] suggests that these websites must be designed in a way that satisfies all users at a local as well as international level. In order to achieve a good level of usability, the website designers and developers should follow design standards and guidelines [30].

The importance of perception of usability and how it can play a role towards the initial impression of a website belonging to a university or a program is emphasised by [44], the role of first impressions of websites on the users was a major factor. Another factor that affects the perception of usability is website architecture which is very crucial for user satisfaction [32], [4]. The designers should try to build a relationship between the students and the university with the help of dialogical exchanges.

Various factors could lead to satisfaction and dissatisfaction pertaining to a particular website; perceived ease of use, perceived usefulness and perceived enjoyment. The users of the Internet are maturing so as their expectations, therefore the service providers should change their approach of designing and functionality [49]. In addition, user expectations change over time due to the rapid growth in technology, this requires the websites to renew their design in order to achieve user satisfaction. The impact of usability guidelines on the aesthetic assessment of e-commerce related

websites and the perceptions of e-retailers is examined [48]; white space, background color, thumbnail image location and size are the four design factors that can change the whole perception of the user. In [50], the author examines the design issues that a multilingual user will have to experience for the user interfaces. A lot of applications are now available in multiple languages. The results show that the English version of the website is much better in terms of usability therefore, the translated versions need improvement.

III. RESEARCH METHODOLOGY

A. University Choices

Six universities were chosen in total, three from the UK; Oxford University (Fig. 2), University of Kent (Fig. 4) and Sheffield Hallam University (Fig. 6) and three from Saudi Arabia; King Abdulaziz University (Fig. 1), Islamic University of Madinah (Fig. 3) and the Arab Open University (Fig. 5).

The reason why UK universities were chosen to make a comparison with Saudi universities is that many UK universities are on top of the world universities rankings table.

The UK university websites are designed in English which is understandable for the Saudi students because the official language of studies in Saudi Arabia is the English language.



Fig. 1. King Abdulaziz University Home Page.



Fig. 2. Oxford University Home Page.

There are many features offered by the university websites, like information related to the courses (description of the course, entry requirement) and services that can help students to get involved with the university (download documents, payment methods, assignment submission, online registration).



Fig. 3. Islamic University of Madinah Home Page.



Fig. 4. University of Kent Home Page.



Fig. 5. Arab Open University Home Page.



Fig. 6. Sheffield Hallam University Home Page.

For this study, six university websites split evenly between the United Kingdom and Saudi Arabia were selected in accordance to their rankings (Table I). For the first pair of universities which this study will compare, Oxford University and King Abdulaziz University were chosen. According to the Times Higher Education World University Rankings, Oxford University is ranked first nationally, similarly, King Abdulaziz University is ranked second nationally. On an international level, King Abdulaziz University is ranked 201-250, while Oxford University is ranked first internationally. The total number of students in King Abdulaziz is more than the total number of students in Oxford University (31554 and 20409 respectively). However, the percentage of international students in Oxford university is higher than the percentage of that in King Abdulaziz University, where Oxford University's percentage of international students is 38%, whereas, the Saudi Arabian university has a total of 21% of international students enrolled.

The second pair of universities that were compared are mid-tier universities, therefore the University of Kent and the Islamic University of Madinah were chosen. However, when collecting statistics on the Islamic University of Madinah, there was no data found on the Times Higher Education World University Rankings website, therefore we had to resort to Webometrics. According to this website, the Saudi Arabian university is ranked 21st nationally, while, Kent University attained a rank of 39 nationally. However, the British university is ranked 301-350 internationally, on the other hand, the Islamic University of Madinah is ranked 5119 internationally. There is a clear and drastic difference between the two universities on an international level even though the Islamic University of Madinah is ranked higher nationally. The total number of students in both universities is similar, as the University of Kent has 164949 students, while the Saudi Arabian university has a total of 20000 students. The Islamic University of Madinah has a very high percentage of international students compared to the University Kent (85% and 31% respectively).

Finally, two universities from the lower tier of the rankings were compared, Sheffield Hallam University and the Arab Open University were selected. As seen previously with the

Islamic University of Madinah, there was no data on the Times Higher Education World University Rankings website for the Arab Open University, therefore we referred to Webometrics. Once again, on a national level, the Saudi Arabian university is ranked higher than the British university. The Arab Open University is ranked 54th nationally and Sheffield Hallam University achieved a rank of 90. Correspondingly, the British university is ranked higher than the Saudi Arabian university internationally. Sheffield Hallam University has been awarded a rank of 801-1000, on the other hand the Arab Open University was 14453 on the international ranking tables on Webometrics. The total number of students as mentioned on the Arab Open University website is 15396. Unlike the previous two pairs of universities compared, where the total number of students was always higher in the Saudi Arabian university, Sheffield Hallam University has more students than the Arab Open University since the British university has 24627 students, of which 14% are international students.

TABLE I. UNIVERSITIES WEB RANKING AND NUMBER OF STUDENTS

University	National Ranking	International Ranking	Number of Students	Percentage of International Students
Oxford University	1	1	20409	38
King Abdulaziz University	2	201-250	31554	21
University of Kent	39	301-350	16949	31
Islamic University of Madinah	21	5119	20000	85
Sheffield Hallam University	90	801-1000	24627	14
Arab Open University	59	14453	15396	n/a

B. Participants

All 60 students who were selected to participate in this study are Saudi male university students aged between 19-25 years, most of them are undergraduate students (85%) and the rest are postgraduate level (15%). To avoid any bias each twenty students evaluated and compared between two websites, one from each country (UK and KSA), each participant spent around twenty minutes to take a tour inside the website.

C. Evaluation Methods

To explore and evaluate usability problems, two methods are used, experiment, user testing [47], [2] and thinking aloud [31]:

1) *Thinking aloud method*: Thinking aloud usability testing method is one of the most effective evaluation methods and most widely used method in usability testing [10], [32] This method supports the user to think aloud and share their thoughts and emotions while carrying out tasks whilst the user is being observed. One of the benefits of this method is that it allows the researcher to comprehend and understand why the user undertakes a specific decision. [27], [28]. Thinking aloud method offers comprehensive details of the information seeking process [42].

2) *Experiment (user testing) method*: User testing is one of the most widely used methods to evaluate website design and to examine the level of usability [47]. Since user testing is the most efficient evaluation method, it is the main method in usability testing [10].

D. Data Collection Process

During the experiment, the participants spoke aloud and then the researcher recorded what was verbalised and took notes during the experiment. Besides that, the participants answered the SUS questionnaire [17] after they spent twenty minutes exploring the website to familiarise with it. Sixty Saudi students compared six websites, one from the UK and the other from KSA, the students were divided into 3 groups each group having 20 students, where they examined and evaluated two websites.

Overall sixty student responses were collected for the comparison between the websites. The SUS questionnaire was selected because it is more accurate, reliable and valid, based on many previous studies [6], [12]-[16], [41].

Moreover, it is the most appropriate method that can be used to compare between different websites (coefficient alpha of .91) [41]. The advantage of SUS is that it can be summarised as being a short list of questions and free to use [39]. The SUS questionnaire contains ten questions, after the participant is familiar with the website (spends twenty minutes before answering certain questions) they are asked to read each question carefully and fill the questionnaire by using the scale (five-point Likert-type scale) from “1” which is strongly disagree to “5” which is strongly agree.

IV. RESULTS

Two powerful methods are used to collect data in this experiment, user testing method and thinking aloud method. The following Tables II, III and IV show the results collected from the students in Saudi Arabian universities.

Table II shows the comparison of usability between Oxford University and King Abdulaziz University. We started our data collection by interviewing 60 students to identify usability flaws in university websites. The students were divided into three groups each group contained 20 students

where they explored usability problems after comparing two websites by completing SUS questionnaire.

The results for King Abdulaziz University website were more varied in comparison to Oxford university website throughout the questionnaire. The results show that 100% of students strongly agreed/agreed that they will continue using Oxford University website frequently as the average score was 4.9, while only 15% agreed to use King Abdulaziz University website frequently with a low score of 2.8.

This partially leads on to the next question where 100% disagreed that the University website was unnecessarily complex with an average score of 1.15, there is a clear correlation between the first two questions, it is expected that if a website is unnecessarily complex, it will disengage the audience and discourage them from revisiting the website, hence the almost perfect correlation between the first two questions in the questionnaire which is why the Saudi University was given a considerably high score of 3.25. In fact, questions two to nine all dictate the results shown in question one. Oxford University achieved a score of 4.75 since most interviewees believed the website was easy to use, whereas King Abdulaziz University scored 3.8 which is respectable. This has a direct link to the next question in the questionnaire, if the students found the website easy to use they would not need the help of an assistant to use the website, therefore Oxford University scored a very low and good score of 1.2 and the Saudi website also scored a good score of 1.55 which is in association to the scores given in the second question.

The fifth question talked about how well the functions were integrated within the website in which Oxford University was given an average score of 4.55, however King Abdulaziz University had a very low score of 2.75. The students believed that there was too much inconsistency in the Saudi website therefore it achieved a score of 3.8, whereas the Oxford University was deemed to be quite consistent as a result attaining a score of 2. Question seven has a direct link to question three, if the user found the website easy to use then they would believe that others would also find it easy to use and vice versa. For Oxford University the scores for question 3 and 7 were almost similar (4.75 and 4.8 respectively), King Abdulaziz University scores were also very similar as it achieved a score of 3.2 (3.3 in question 3), this proves the clear and distinct correlation.

Question eight discusses how cumbersome the website was to use, Oxford University was given an average score of 1.1 which is perfect and the Saudi website had a score of 1.7. Most students felt very confident using the Oxford University website (score of 4.3) and the number of students who did not feel confident using King Abdulaziz website had a score of 2.85. The final question involves how quickly the user got accustomed to the website, when using the Saudi website, most students got accustomed to the website quickly and this gave King Abdulaziz University a score of 2.4 on this question, the Oxford University website was easier hence they achieved a better score of 1.35.

The average overall SUS Score indicates how successful a website is (in this case the two university websites chosen) in

which the Oxford University website gained an exceptional score of 91.0 (grade A+) whereas the King Abdulaziz University website gained a poor score of 55.5 which means that the Saudi university website requires intense adjustments in order to achieve a higher score.

Table III shows the comparison between the University Kent's website and the Islamic University of Madinah's website. The second set of 20 students gave their opinions on the university websites via filling out the SUS questionnaire.

The first question discusses whether or not the user would revisit the website, the University of Kent achieved an average score of 4.8 which is close to perfect, whereas the Islamic University of Madinah achieved an average score of 3.3 which on the other hand is respectable. Question two talks about the simplicity of the website where 1 on the scale means perfection since the questions states "I found this website to be unnecessarily complex", therefore 5 on the scale would be the worst score that can be given; the University of Kent scored 1.9 and the Islamic University of Madinah scored 2.5 which is quite close to the score awarded to the University of Kent, this data is normally distributed as most of the student answers were spread out from 1-2 and 4-5, but compact in the middle as half of the students awarded the University of Kent a score of 3.

Question three questions how easy it was for the user to navigate through the website, the Saudi university website had been awarded a score of 3.1 whereas on the other hand the British university was given a score of 4.35 as most students believed that it was easy to use the website. Question four relates somewhat to the previous question, if the website was easy to use there would be no need for an assistant which is exactly what this question is asking ("I will need the help of a support person"), hence we would expect the results to be almost simultaneous; since Kent university was given a high score in question three it is expected to have a low score on question four, this university achieved a score of 1.1 where only 10% of the students gave it a score of 2, unlike the Islamic University of Madinah where the results were more spread out and the average score was 3.

For question five, the students gave the Saudi Arabian website a score of 2.2 however they awarded the University of Kent a score of 4.7 which is exceptional. The level of consistency is the topic of the next question, where the students provided the Islamic University of Madinah with a score of 3.7 which needs to be improved on, on the other hand the University of Kent was assigned with a score of 1.5 which is remarkable.

For the next question the students are asked if they think other people would learn to use this website very quickly which is very subjective and would be based on whether or not the students found the website easy to use themselves, therefore we would expect the results from question three and this question to be very similar. The University of Kent received a score of 4.8 (in question three 4.35) and the Islamic University of Madinah received a score of 2.9 and in question three they attained a score of 3.1. In question seven, the students were asked about how cumbersome they found the website to be, most students gave the University of Kent a

score of 1 which gave an average score of 1.2, however the Saudi Arabian university was awarded a score of 3.1. The level of confidence was examined in the next question, students felt much more confident using the University of Kent website than the Islamic University of Madinah (a score of 4.7 and 2.7 respectively).

For the final question, almost all students strongly disagreed with this statement with regards to the University of Kent website, thus receiving an average score of 1.25, conversely, the Islamic University of Madinah attained a score of 3.35. These ten questions are combined and give an average overall SUS score which proves how well a website is being run. In this case the Islamic University of Madinah had an overall score of 46.4 which on the Curved Grading Scale for the SUS is in the lowest tier (Grade F), [11].

Table IV of this experiment includes Sheffield Hallam University and the Arab Open University. For the first question which states “I would like to use this website frequently”, the Sheffield Hallam University achieved a score of 3.5, whereas the Arab Open University achieved a similar score of 3.1. In the second question it discussed how complex the participants found the website to be, Sheffield Hallam University received a score of 2.05 unlike the Arab Open University which achieved a less satisfying score of 2.25 (for this question the higher the mark the lower the total SUS Score).

Question three discusses the ease of use regarding the website from the student’s perspective, results show that students believed that the Sheffield Hallam University website was easier to use hence a score of 4.15 was awarded; therefore the Arab Open University attained a lower score of 3.9. Question four says “I would need the help of a support person” which means that the participants struggled to use this website, Sheffield Hallam University attained a score of 1.35, while the Saudi university achieved a score 1.8 which is also respectable. Sheffield Hallam University achieved a high score of 4.1, whereas the Arab Open University achieved a poor score of 2.45.

Question six discusses the level of consistency, where Sheffield Hallam University was thought to be more consistent than the Arab Open University achieving scores of 1.8 and 3.95 respectively (for this question the higher the mark the lower the total SUS Score). For question seven, participants were asked to comment on if they believed other people will adapt to this website very quickly or not, for the Sheffield Hallam University website most people believed that it was quite easy to understand and adapt to, hence receiving a score of 3.9, however the Arab Open University wasn’t as easy to learn to use, as a result receiving a marginally lower score of 3.6.

Question eight enquires about how cumbersome the website was to use, Sheffield Hallam University achieved a score of 1.5 whereas the Arab Open University achieved a score of 2.1 (for this question the higher the mark the lower the total SUS Score) where both scores are respectable. For question nine, participants were asked to comment on how

confident they were when using the website, both universities achieved the exact same score of 2.6. In the final question, students were asked to discuss if they needed to learn a lot of things before they could get going with the website; Sheffield Hallam University achieved a score of 2.2 while the Arab Open University received a score of 2.6.

These ten scores were used to calculate an overall average SUS score in which Sheffield Hallam received an overall score of 69.3, whereas the Arab Open University achieved an overall SUS score of 61.5. Table V shows universities scores based on SUS questionnaire scales and total SUS scored by each University

TABLE II. COMPARISON BETWEEN OXFORD UNIVERSITY (GREEN) AND KING ABDULAZIZ UNIVERSITY (ORANGE)

The SUS Questionnaire					
	Strongly Disagree (1)	2	3	4	Strongly Agree (5)
1. I think that I would like to use this system frequently.				10	90
	5	30	45	20	
2. I found the website to be unnecessarily complex.	85	15			
		20	40	35	5
3. I thought the website was easy to use.		5		10	85
	5	25	20	35	15
4. I think that I would need the support of a technical person to be able to use this system.	80	20			
	55	35	10		
5. I found the various functions in this website were well integrated.		5	5	20	70
	10	30	35	25	
6. I thought there was too much inconsistency in this website.	10	80	10		
		15	20	35	30
7. I would imagine that most people would learn to use this system very quickly.			5	10	85
	10	30	45	15	
8. I found the system very cumbersome to use.	90	10			
	45	40	15		
9. I felt very confident using the system.	80	10	10		
	10	25	35	30	
10. I needed to learn a lot of things before I could get going with this system.	75	15	10		
	15	35	45	5	

TABLE III. COMPARISON BETWEEN THE UNIVERSITY OF KENT (WHITE) AND ISLAMIC UNIVERSITY OF MADINAH (BLUE)

The SUS Questionnaire					
	Strongly Disagree (1)	2	3	4	Strongly Agree (5)
1. I think that I would like to use this system frequently.			5	10	85
		20	45	20	15
2. I found the system unnecessarily complex.	35	40	25		
	10	35	50	5	
3. I thought the system was easy to use.			5	55	40
		20	35	30	15
4. I think that I would need the support of a technical person to be able to use this system.	90	10			
		25	55	15	5
5. I found the various functions in this system were well integrated.				30	70
	20	55	10	15	
6. I thought there was too much inconsistency in this system.	55	40	5		
			55	20	25
7. I would imagine that most people would learn to use this system very quickly.				20	80
	20	25	20	15	20
8. I found the system very cumbersome to use.	80	20			
9. I felt very confident using the system.			5	20	75
	25	25	30	15	10
10. I needed to learn a lot of things before I could get going with this system.	85	15			
		20	25	55	

TABLE IV. COMPARISON BETWEEN SHEFFIELD HALLAM UNIVERSITY (GREY) AND ARAB OPEN UNIVERSITY (YELLOW)

The SUS Questionnaire					
	Strongly Disagree (1)	2	3	4	Strongly Agree (5)
1. I think that I would like to use this system frequently.		10	40	40	10
		20	50	30	
2. I found the system unnecessarily complex.	35	30	30	5	
	20	35	45		
3. I thought the system was easy to use.			15	55	30
			35	40	25
4. I think that I would need the support of a technical person to be able to use this system.	70	25	5		
	45	30	25		
5. I found the various functions in this system were well integrated.		10	10	40	40
	5	55	30	10	
6. I thought there was too much inconsistency in this system.	40	40	20		
			35	35	30
7. I would imagine that most people would learn to use this system very quickly.			30	50	20
		10	40	30	20
8. I found the system very cumbersome to use.	55	40	5		
	35	30	25	10	
9. I felt very confident using the system.	25	30	10	35	
	10	25	55	10	
10. I needed to learn a lot of things before I could get going with this system.	20	40	40		
	10	30	50	10	

TABLE V. UNIVERSITY SCORES BASED ON SUS QUESTIONNAIRE SCALES AND TOTAL SUS SCORE OF EACH UNIVERSITY

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SUS Score
Oxford University	4.9	1.15	4.75	1.3	4.55	2	4.8	1.9	4.3	1.35	91
King Abdulaziz University	2.8	3.25	3.3	1.55	2.75	3.8	3.2	1.7	2.85	2.4	52.9
University of Kent	4.8	1.9	4.35	1.1	4.7	1.5	4.8	1.2	4.7	1.25	90.8
Islamic University of Madinah	3.3	2.5	3.1	3	2.2	3.7	2.9	3.1	2.7	3.35	46.4
Sheffield Hallam University	3.5	2.05	4.15	1.35	4.1	1.8	3.9	1.5	2.6	2.2	73.4
Arab Open University	3.1	2.25	3.9	1.8	2.45	3.95	3.6	2.1	2.6	2.6	57.4

V. DISCUSSION

The following section will discuss the results found in the previous section, where Arabic university websites are going to be compared to British university websites.

Usability is an important factor to evaluate and measure different websites. Moreover, the quality of designing a website plays a major role in the usability factor [9], [34]. This study used the SUS questionnaire and thinking aloud method as the main resources to analyse the level of usability between two sites. The idea behind using SUS questionnaire is that it measures how good or bad the website is and the level of satisfaction regarding the users [13], [14]. The average score generally is 68, anything below suggests there are serious problems with the website usability. However, if it is above 68 then the satisfaction of users when discovering the website is high. [8], [38]. In order to get a grade A, a score of 80.3 or above is required which may lead to users recommending the website to others. 68 is graded as C and 51 is graded F indicating serious problems with the site, therefore an urgent review of the website must be made.

The aim of this study is to explore usability problems within Saudi university websites in comparison with the UK university websites by applying the SUS methods and 'Thinking aloud method' to measure the usability of the websites. The results of this study illustrated that students of the Saudi university faced various usability problems throughout the experiment. The most common usability problems found by students were satisfaction, integration and confidence regarding King Abdulaziz University website.

The satisfaction of students directly affects their confidence, for that reason they are not willing to use the website frequently. This consequently affected the confidence of students when using the website. Besides that, the results also showed that various functions in the King Abdulaziz University website were not integrated well. This is evident from the website where the text, icons and images are not fully integrated. Fig. 7 shows the bad integration of the website, for example when a drop icon is selected a black line appears instead of drop menu. Another example of bad integration in the website is shown in Fig. 8 where the writing clearly. The student can use the website without the need for help and this consequently affects the ease of use of the website in a positive way i.e. ease of use does not mean the usability of the website is efficient and effective [36], [46].



Fig. 7. Black Line Shows Poor Integration.



Fig. 8. Writing Overlaps.

Overall Oxford University scored 91.0 which means most students were satisfied to use the website without facing major usability problems, while the average score of King Abdulaziz University was 52.9 which is grade D which shows there are serious problems with the website. This indicates that the King Abdulaziz University website requires urgent review and may need to be redesigned to achieve user satisfaction. King Abdulaziz University can utilise the Oxford University website and use it as a base to improve its own website since it scored a very good score (91.0).

Moving on, the University of Kent had an overall score of 90.8 which is an outstanding score of A+ (the highest score that can be achieved on this grading scale) compared to that of Islamic University which scored 46.4 (grade F). This indicates that most of the students were not satisfied with the Islamic University website based on the average score, consequently meaning the website requires a major overhaul. For that reason, since the University of Kent scored very high, the Islamic University website can use that as a base when redesigning their website.

The results obtained indicates there are various usability problems within the Islamic University website. These problems were identified as integration, inconsistency of the website, learning how to use the website quickly and finally confidence when using the website. Integration and inconsistency are mutually inclusive, this is evident from the student's response. The website contained plenty of examples of how integration is poor and there is no consistency in the majority of pages of the website, as an example, when the students were asked to observe the website before filling the questionnaire they found many problems within a short time, for example when some of them wanted to search about some courses they discovered there is no search engine to help them navigate through the website (Fig. 11), secondly they could not find any information about the majority of courses in different faculties, the only information that was available is some announcements in a few faculties which are not updated. In addition, the Arabic writing was aligned from left to right, when instead it should be right to left (Fig. 10). It was also difficult to find any information when navigating within the faculty pages. The poor quality of images (Fig. 9), within the website in general and the structure of the websites was very bad based on student's response this included icons, images, text, language and font type.



Fig. 9. Image Being Stretched, Lowering Quality.



Fig. 10. Arabic Writing should be from Right to Left.



Fig. 11. English Page Shows Arabic Writing, No Search Engine.

By using the ‘Thinking aloud method’ students identified certain problems with the website which are shown in Fig. 12 where the page wasn’t fully translated to English. Since the integration of the website was below standard, this led to certain aspects of the website design being poor. In addition to that students found that it was difficult to learn how to use the website quickly. One of the reasons which may contribute to this factor is that the English website is not fully translated from Arabic to English which may cause confusion amongst non-Arabic speakers. Another apparent problem was that after selecting an option from the menu bar at the home page there is no option available to help navigate through the website (i.e. changing from one faculty to another) as you are forced to go back to the home page and select from the menu bar again. The majority of the students did not feel comfortable using the website due to the lack of integration, inconsistency in translation and difficulty to use the website easily.

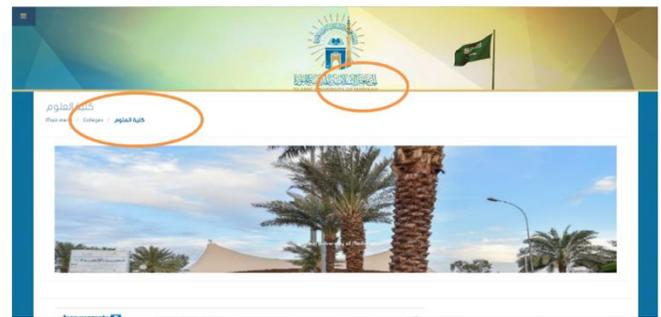


Fig. 12. Once Again Arabic Writing on English Page.

In contrast the University of Kent website gained high scores and attracted students based on various points, the main ones being very quick to learn regarding how to use the website followed by the ease of navigation, good integration of site and consistency of the website. These are the main factors that influenced the confidence and the satisfaction of Saudi Arabian students whom used the website.

The comparison between websites was chosen to evaluate the lower ranked universities in the UK as well as Saudi Arabia. Sheffield Hallam University and the Open Arab University were examined during this study. The total score for Sheffield Hallam University was 73.4 (grade B-) whilst the Arab Open University was 57.4 (grade D). Sheffield Hallam University website requires some attention to improve the usability of the website such as satisfaction of students, for example when the ‘Study here’ option on the navigation bar is selected various options pop up. When ‘find a course’ is selected, another ‘find a course’ will appear again (Fig. 13) which confused the students since the same terminology was used twice without them being differentiated. To solve this, it would be better if the colour of the subtitle was changed or one of the two icons was renamed.

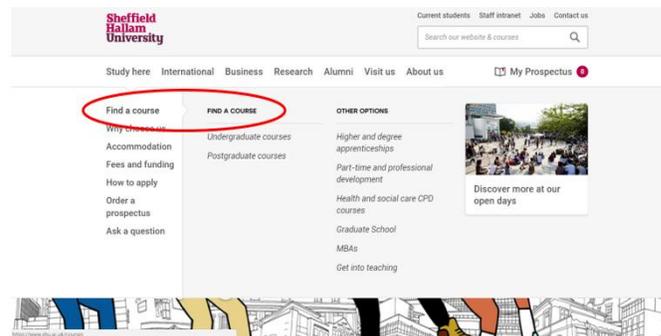


Fig. 13. ‘Find A Course’ Option Shown Multiple Times.

Based on the comments of the students, it was clear that their confidence had decreased. This was observed via ‘Thinking aloud method’. The two main problems that were identified within the Arab Open University website included inconsistency and integration. The main function of any university website is to offer enough information with content about the types of facilities available, the type of courses available (graduate and undergraduate), fees and funding, international/national applications and student life. This is not the case for the Arab Open University as the students found that within the faculty page detailed information about each

faculty and its departments (i.e. courses) was not available as shown in Fig. 14. This ultimately affects the inconsistency of the websites, because of this, the students did not feel comfortable nor satisfied when using the website.

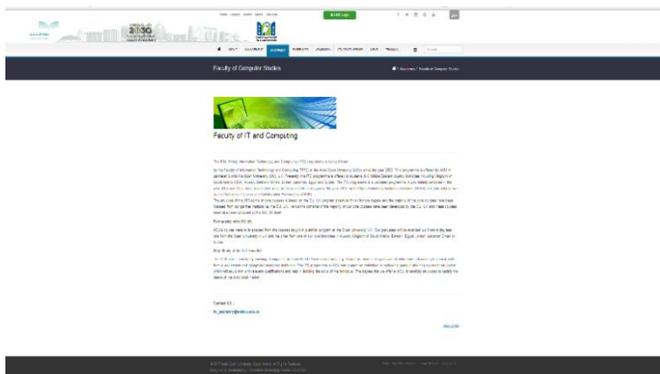


Fig. 14. Insufficient Information about the Course, Only Mentions Faculty.

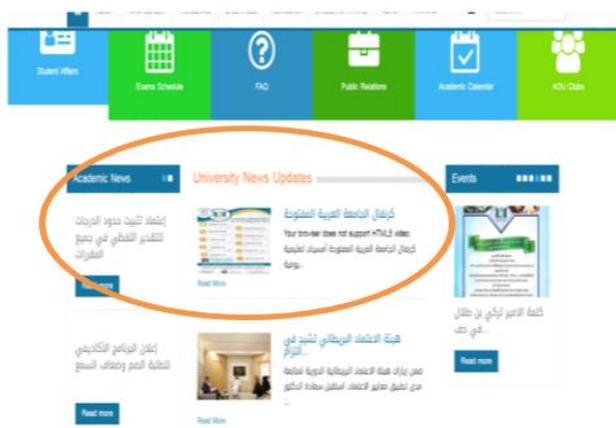


Fig. 15. Webpage Not Fully Translated.

Besides that, the translation of the content is not fully translated from Arabic to English (Fig. 15). Even the Arabic content available in certain pages are aligned in English format. Another problem which arose is the integration of the website, an example that was detected by the students was that some icons were useless as they did not direct you back to the link selected i.e. when 'academics' is selected the hyperlink does not work, likewise the home logo has no hyperlink to direct you to the home page.

VI. CONCLUSION

The main goal of this study is to evaluate the Arabic university websites in comparison with the UK university websites. Six websites from three different university tier rankings (high, mid and low) were examined to demonstrate if there are any usability differences between the three levels. These levels were chosen to show if the different levels of university ranking influenced the usability of the website. The results showed that all Saudi university websites within the three levels (high, mid and low) shared three major problems. The problems that arose during this study were inconsistency, integration and satisfaction factors.

The results that were collated proved that the Arabic university websites faced numerous problems including satisfaction, integration, inconsistency, confidence, design concepts, ease use of websites and comfort of students. These major usability problems effect the Saudi websites directly and need to be solved urgently to attract the students and enable them to benefit from the website. Problems like mixed language, poor content, lack of information, satisfaction and icons can harm the website leading to students not interacting with the website and in the opposite way attracting content is the key feature of any effective website [33], [7], [9].

VII. LIMITATIONS OF STUDY

Based on the number of university websites that were evaluated in this study, the quantity may limit the findings and the number of students that evaluated the websites may not be of a sufficient number to generalise the finding of this study. Besides that, the participants involved in this study were only male which may limit the results found, if male as well as females were part of the experiment it will ultimately increase the number of participants which validates the findings.

VIII. FUTURE RECOMMENDATIONS

For future research, additional usability problems within Arabic university websites should be investigated, this may include further testing of websites from different countries as well as a large number of students which includes male and females to improve the validity and the efficiency of Arabic university websites. The satisfaction of users can be reached if the designers of websites give more attention to the usability problem in order to design high quality websites.

REFERENCES

- [1] C.Flavian, R. Gurrea, and C. Orús, "Web design: a key factor for the website success," Journal of Systems and Information Technology, vol. 11, no. 2, pp. 168-184, 2009.
- [2] E. Liljegren, and A. L. Osvolder, "Cognitive engineering methods as usability evaluation tools for medical equipment," International Journal of Industrial Ergonomics, vol. 34, no. 1, pp. 49-62, 2004.
- [3] E. Şengel, "Usability Level of a University Web Site," Procedia - Social and Behavioral Sciences, vol. 106, pp. 3246-3252, 2013.
- [4] G. Shivaprasad, N. S. Reddy, U. D. Acharya, and P. K. Aithal, "Neuro-Fuzzy Based Hybrid Model for Web Usage Mining," Procedia Computer Science, vol. 54, pp. 327-334, 2015.
- [5] ISO, "international standard organisation, ISO 9241 – 11 – 03 – 15, part 11," guidance on usability 1st Edition, 1998.
- [6] J. Brooke, "SUS-A quick and dirty usability scale," Usability evaluation in industry, vol. 189, no. 194, pp. 4-7, 1996.
- [7] J. Hartmann, A. Sutcliffe, and A. De Angeli, "Investigating attractiveness in web user interfaces," In Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, pp. 387-396, 2007.
- [8] J. J. R. Lewis, and J. Sauro, "Revisiting the Factor Structure of the System Usability Scale," Journal of Usability Studies, vol. 12, no. 4, 2017.
- [9] J. Melorose, R. Perroy, and S. Careas, "Usability Measurement: A Roadmap for a Consolidated Model," Stew. Agric. L. Use Baseline, vol. 1, 2015.
- [10] J. Nielsen, and H. Loranger, "Prioritizing web usability," Pearson Education.,2006
- [11] J. R. Lewis, and J. Sauro, "Item Benchmarks for the System Usability Scale," Journal of Usability Studies, vol. 13, no. 3, 2018.

- [12] J. R. Lewis, and J. Sauro, "The factor structure of the System Usability Scale," In Kurosu, M. (Ed.), *Human Centered Design, HCI, Heidelberg, Germany: Springer-Verlag*, pp. 94–103, 2009.
- [13] J. R. Lewis, B. S. Utesch, and D. E. Maher, "Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability," *International Journal of Human-Computer Interaction*, vol. 31, pp. 496–505, 2015.
- [14] J. R. Lewis, B. S. Utesch, and D. E. Maher, "UMUX-LITE – When there's no time for the SUS," In *Proceedings of CHI, Paris, France: ACM*, pp. 2099–2102, 2013.
- [15] J. R. Lewis, J. Brown, and D. K. Mayes, "Psychometric evaluation of the EMO and the SUS in the context of a large-sample unmoderated usability study," *International Journal of Human-Computer Interaction*, vol. 31, no. 8, pp. 545–553, 2015.
- [16] J. Sauro, "A practical guide to the system usability scale: Background, benchmarks, & best practices," Denver, CO: *Measuring Usability LLC*, 2011.
- [17] K. Finstad, "The system usability scale and non-native English speakers," *Journal of Usability Studies*, vol. 1, no. 4, pp. 185-188, 2006.
- [18] L. Hasan, "Can Students Complete Typical Tasks on University Websites Successfully?" In *5th International Conference on Education and Educational Technologies*, pp. 141-147, 2014.
- [19] L. Hasan, "Evaluating the usability of nine Jordanian university websites," In *Communications and Information Technology (ICCIT), International Conference on IEEE*, pp. 91-96, 2012.
- [20] L. Hasan, "The Website of the University of Jordan: Usability Evaluation," *Int. Arab J. e-Technology*, vol. 3, no. 4, pp. 258-269, 2014.
- [21] L. V. Casalo, C. Flavián, and M. Guinaliú, "The role of satisfaction and website usability in developing customer loyalty and positive word-of-mouth in the e-banking services," *International Journal of Bank Marketing*, vol. 26, no. 6, pp. 399-417, 2008.
- [22] M. A. Ababtain and A. R. Khan, "Towards a Framework for Usability of Arabic-English Websites," *Procedia Computer Science*, vol. 109, pp. 1010-1015, 2017.
- [23] M. Aliyu, M. Mahmud, and A. O. M. Tap, "Exploring Islamic website features that influence user satisfaction: A conceptual model," *Procedia-Social and Behavioral Sciences*, vol. 65, pp. 656-661, 2012.
- [24] M. Benaida, and A. Namoun, "Technical and Perceived Usability Issues in Arabic Educational Websites," *International Journal of Advanced Computer Science and Applications(IJACSA)*, Vol. 9 no. 5, 2018.
- [25] M. H. Thowfeek and M. N. Abdul Salam, "Students' Assessment on the Usability of E-learning Websites," *Procedia - Social and Behavioral Sciences*, vol. 141, pp. 916-922, 2014.
- [26] M. Hamiti, B. Reka, and F. Imeri, "The Impact of Computer Components in Enhancing the Quality of Teaching and Learning Process in Universities," *Procedia - Social and Behavioral Sciences*, vol. 191, pp. 2422-2426, 2015, 2015.
- [27] M. Koro-Ljungberg, E. P. Douglas, D. Therriault, Z. Malcolm, and N. McNeill, "Reconceptualizing and decentering think-aloud methodology in qualitative research," *Qualitative Research*, vol. 13, no. 6, pp. 735–753, 2013.
- [28] M. L. Frigotto, "Effectuation and the Think-Aloud Method for Investigating Entrepreneurial Decision Making," In *Complexity in Entrepreneurship, Innovation and Technology Research*, pp. 183-197, 2016.
- [29] M. M. Elobaid, A. Lodhi and A. R. Khan, "Usability Testing of Multilingual Educational Websites," *Life Science Journal*, vol. 10, no. 2, 2013.
- [30] M. Masood and A. Musman, "The Usability and its Influence of an e-Learning System on Student Participation," *Procedia - Social and Behavioral Sciences*, vol. 197, pp. 2325-2330, 2015.
- [31] M. T. Boren, and J. Ramey, "Thinking aloud: reconciling theory and practice," *IEEE Trans. Professional Communication*, vol. 43, pp. 261 – 278, 2000.
- [32] M. Zviran, C. Glezer, and I. Avni, "User satisfaction from commercial web sites: The effect of design and use," *Information & management*, vol. 43, no. 2, pp. 157-178, 2006.
- [33] P. J. Lynch, and S. Horton, "Web Style Guidelines (2nd ed.)," Yale University Press, New Haven, CT, 2001.
- [34] P. Yu, and D. Zhao, "Effect of website quality factors on the success of agricultural products B2C e-commerce," *International Conference on Computer and Computing Technologies in Agriculture Springer, Berlin, Heidelberg*, pp. 98-113, 2013.
- [35] P. Zhang, and G. von Dran, "User expectations and rankings of quality factors in different Web site domains," *International Journal of Electronic Commerce*, vol. 6, no. 2, pp. 9-33, 2002.
- [36] Q. Whitney, "Balancing the 5Es: Usability," *Cutter IT Journal*, vol. 17, no. 2, 2004.
- [37] R. Redekopp, and K. Kalanda, "Internet use: a study of preservice education students in Lesotho and Canada," *Procedia-Social and Behavioral Sciences*, vol. 182, pp. 529-534, 2015.
- [38] S. B. Linek, "Order Effects in Usability Questionnaires," *Journal of Usability Studies*, vol. 12, no. 4, 2017.
- [39] S. Borsci, S. Federici, S. Bacci, M. Gnaldi, and F. Bartolucci, "Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, UMUX-LITE as a Function of Product Experience," *International Journal of Human-Computer Interaction*, Taylor & Francis Group, LLC, vol. 31, pp. 484–495, 2015.
- [40] S. Bozyigit, and E. Akkan, "Linking Universities to the Target Market via Web Sites: A Content Analysis of Turkish Private Universities' Web Sites," *Procedia - Social and Behavioral Sciences*, vol. 148, pp. 486-493, 2014.
- [41] S. C. Peres, T. Pham, and R. Phillips, "Validation of the System Usability Scale (SUS) SUS in the Wild," In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Sage CA: Los Angeles, CA: SAGE Publications, Vol. 57, No. 1, pp. 192-196, 2013.
- [42] T. K. Hoppmann, "Examining the 'point of frustration'," *The think-aloud method applied to online search tasks, Quality & Quantity*, vol. 43, no. 2, pp. 211-224, 2009.
- [43] T. Ramayah, N. H. Ahmad, and M. C. Lo, "The role of quality factors in intention to continue using an e-learning system in Malaysia," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 5422-5426, 2010.
- [44] T. Sundeen, K. M. Vince Garland, and W. D. Wienke, "A Multi-Year Evaluation of Student Perceptions of University and Special Education Doctoral Websites," *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, vol. 39, no. 4, pp. 259-275, 2016.
- [45] U. Sharkey, M. Scott, and T. Acton, "The influence of quality on e-commerce success: an empirical application of the Delone and Mclean IS success model," *International Journal of E-Business Research (IJEER)*, vol. 6, no. 1, pp. 68–84, 2010.
- [46] W. Quesenbery, "What does usability mean: Looking beyond ease of use," In *Annual conference-society for technical communication*, vol. 48, pp. 432-436, 2001.
- [47] W. S. Tan, D. Liu, and R. Bishu, "Web evaluation: Heuristic evaluation vs. user testing," *International Journal of Industrial Ergonomics*, vol. 39, no. 4, pp. 621-627, 2009.
- [48] A. Agarwal and A. Hedge, "The Impact of Web Page Usability Guideline Implementation on Aesthetics and Perceptions of the E-Retailer," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2008.
- [49] C.S. Ong, S.C. Chang and S.M. Lee, "Website satisfaction dimension: Factors between satisfaction and dissatisfaction," *Information Development Journal*, 2013.
- [50] M. Mahdi and A. Maaruf, "User interface (UI) design issues for multilingual users: a case study," *Universal Access in the Information Society*, 2016.

Using Sab-Iomha for an Alpha Channel based Image Forgery Detection

Muhammad Shahid Bhatti¹, Syed Asad Hussain², Abdul Qayyum³,
Abdul Karim Shahid⁴, Muhammad Usman Akram⁵, Sajid Ibrahim Hashmi⁶
Department of Computer Science,
COMSATS University Islamabad,
Lahore Campus, Pakistan

Abstract—Digital images are a very popular way of transferring media. However, their integrity remains challenging because these images can easily be manipulated with the help of software tools and such manipulations cannot be verified through a naked-eye. Although there exist some techniques to validate digital images, but in practice, it is not a trivial task as the existing approaches to forgery detection are not very effective. Therefore, there is need for a simple and efficient solution for the challenge. On the other hand, digital image steganography is the concealing of a message within an image file. The secret message can be retrieved afterwards by the author to check the image file for its veracity. This research paper proposes Sabiomha, an image forgery technique that make use of image steganography. The proposed technique is also supported by a software tool to demonstrate its usefulness. Sabiomha works by inserting an invisible watermark to certain alpha bits of the image file. The watermark we have used to steganograph an image is composed of a combination of text inputs the author can use to sign the image. Any attempts to tamper the image would distort the sequence of the bits of the image pixel. Hence, the proposed technique can easily validate originality of a digital image by exposing any tampering. The usability of our contribution is demonstrated by using the software tool we developed to automate the proposed technique. The experiment which we performed to further validate our technique suggested that Sabiomha could be flawlessly applied to image files.

Keywords—Digital images; tamper; steganography; metadata; forgery detection; cipher; image authentication; image validation; watermarking

I. INTRODUCTION

A. Background

Applications of digital images have been the focal point of computer vision researchers for decades now [1]–[4]. Digital content is used as an effective way of communication among different stakeholders [5]. The advent of digital devices and communication technologies has led to increase in the use of image files for sharing visual moments and photographs. Digital images are generated through cameras with-out transformation and development process contrary to camera reels in the past and can be delivered electronically through any supporting communication channel.

Although an image data is generally considered reliable but with the passage of time, the digital technology itself has compromised the faith we have had in electronic content. The ever-increasing trend of malpractices in image forensics has posed new challenges to the research horizon as we continue

to exist in the era which is very much vulnerable to multiple facets of digital contents. The situation seeks effective and efficient solution to ensure integrity of digital images.

With multi-million users using emails and social media, nearly countless digital content is distributed and shared every day. A large portion of the content comprises of digital images. These days users can easily capture their memorable moments through digital cameras and can share with others by publishing the image files on the web. On the other hand, users can potentially receive tampered images and unknowingly circulate those as well. Since digital data is easily accessible these days, obnoxious users can manipulate image files for entertainment and at times abuse those for some societal or political gains or to dictate any legal affairs. This phenomenon is reinforced by the availability of some supporting software applications. Hence the situation calls for taking some concrete measures to meet these challenges.

Previously, digital forensics domain has helped to rejuvenate some trust in digital content. However, as the image forgery detection techniques are being developed, tampering of digital data despite leaving any noticeable trails has become very trivial. The challenge leads to issues such as image authentication, protection, and forgery detection. This demands aggressive counter approaches from scientists and researchers to confront and challenge malpractices.

B. Problem Description

Image tampering is a known handling technique [5]. Deception of typical image files is relatively a tedious task and requires sufficient expertise. However, digital images are disposed to tinkering. There exist numerous software applications to easily manipulate them. Malpractices mainly include duplication, replication, removing or exchanging parts of an image. It should be noted that originality of an analog data can be validated easily through a naked eye as any attempts to tampering can be conceived readily. Contrarily, development of supporting software tools has made manipulation of digital images a very easy task. For example, Fig. 1 highlights one such example. Originally, two objects were present in Fig. 1(a). The object on the far right is inserted as visible in Fig. 1(b). However, by looking at the figure through a naked eye, one cannot conceive that originality of the image had been compromised. Before taking an appropriate legal or social action in such cases, it is necessary to verify that an image had been edited. In such cases, as it is clear from the figure, validation of originality of a tampered image becomes very

challenging since alteration of a digital image can be carried out easily in comparison to a printed one.

As digital image domain is being revolutionized, tampering of a digital content without any noticeable impression has become very effortless. Therefore, to tackle the challenge, an image should be analyzed in such a way that even a slight attempt to forge can be detected straightaway. In this paper, we propose a light-weight automated technique that image owners and publishers can easily use to sign their images. The approach can also be used as an instrument to protect proprietary images from any possible forgery attempts.

Rest of the paper is organized as follows: following subsections of Section 1 highlight the contribution and the current state of the art in the domain. Section 2 describes the related work. Section 3 reflects upon our contribution in terms of the proposed technique and presents its usefulness through a software tool we developed to automate and demonstrate our work. In the end, Sections 4, 5 and 6 sum up with Automation of SAB - IOMHA, and Conclusions, respectively.

C. Contribution of the Research Project

Validation is a standard procedure for investigating integrity of an object. We want to achieve it in terms of forgery detection of a digital image through the proposed work. The decisive objective is to audit digital image files for originality and verify that their integrity has not been compromised since their authoring. The current approaches for the purpose have encompassed signature-based methods for protecting image files and checking for their integrity. However, such techniques are not applicable in wider settings because of their limitations or overheads involved in their use. On the other hand, as part of our work, we propose using a composite watermark which consists of a cipher along with date and time stamp and email address of the image author. The watermark is inserted in structured patterns to certain bits of an image file.

Digital watermarking is a known technique for media files for retaining copy-right information and identification of their proprietorship [4]. These can be of several types and are widely used. Generally, images can be inserted with at least two types of watermarks, visible watermarks or invisible watermarks as required. A visible watermark embeds an image file with an identification mark and an invisible one on the other hand inflicts a hidden mark in it. As part of this research, we choose the invisible watermark which we sequentially insert across multiple bits of a digital image. The contents and structure of the watermark is distorted if someone tries to edit the image file by any means.

In this research paper we provide more insight and extend Sab-iomha which we proposed previously [6], for its usefulness in the real settings. The extended version of the work reflects upon more technicalities of the technique and an improved validation mechanism. The ultimate objective of the research is to address the challenge of digital image forgeries.

D. Current State of the Art

ELA (Error Level Analysis) of an image can highlight any edited or distorted part of an image as different regions of an image having different compression levels can be identified.

It enables the stakeholders to easily detect any problem areas through a naked-eye. Existing approaches to image forgery detection usually involve replicating those files to some dedicated software tools [7]. Users are then provided with different features of ELA and Joint Photographic Expert Group (JPEG) format. Our contribution is twofold. First, we split an image file to temporarily separate its metadata from the visual content and then steganograph the same image. An image file is composed of combination of pixels. An ordered set of bytes represents each pixel for different colors that constitute an image. Those colors include Alpha, Red, Green, and Blue. It should be noted that data is not stored in Alpha bits. Therefore, we propose use of those bits to insert the hidden watermark into the image file. Cipher, as part of the watermark, is invisible and is removed automatically upon any attempts to forge. As part of the second contribution of the research paper, we demonstrate the usefulness of Sab-iomha through automation in terms of a software tool we developed to augment the proposed technique. If an image file was saved multiple times, it loses its quality [8]. Metadata of an image file refers to the image itself. The information it contains may include the image type; e.g. JPEG, dimensions of the image, internal formats, and color scheme. The metadata also gives information such as the date of creation, the date of modification, name of the software editor that was used to create the image, file tags, and camera tags. It also provides information on the Exchangeable Image File Format (EXIF) which is used by the digital cameras manufacturers to extract camera settings that were used to capture the image. Camera settings entail information such as the manufacturer name and the model, time stamp, and lens settings. Those settings may vary among images to ensure maximum level of integrity. If a user tries to insert comments into an image file, they are incorporated into its metadata. Digital cameras normally do not allow automatic insertion of comments to the captured image. However, if any additions are found, it is an indication that the image has been edited or reprocessed using some software tool.

Majority of the existing approaches to image forgery detection take account of the information provided through metadata or the file header. Any attempts to get additional information while capturing a digital photo or any effort to change its header can easily render image handling more complex hence time consuming. In addition to that, the currently available techniques do not account for digital contents or file storage itself. On the other hand, our proposed technique addresses the challenge using a simple yet efficient mechanism; i.e. hidden watermark is embedded in an image which diminishes the need for manipulating with the file header. Sab-iomha ensures that any attempts to manipulate the image distort the watermark. Hence any successful bids to alter the image file can be discovered promptly.

II. RELATED WORK

The literature review that was conducted to carry out this research encompassed image content, detection, and forensic analysis. We investigated different techniques currently in use for authenticating digital contents in terms of their traits as well as deficiencies.

Lighting, inconsistent shading, and shadows have been used as a method for collecting evidence on image forgery [5].



(a)



(b)

Fig. 1. (a) Original image (b) Object on the right is inserted.

Mixture of shadow and shading was rationally used to serve for the purpose and both were made dependent on each other but in case they are not, the corresponding image is found to be a tampered one. Furthermore, the authors reported reliable and specific shadings under different inferences of some subjective measures such as guess-work or acceptance. However, their proposed technique is not applicable in case such historical text documents do not make a shadow. Moreover, the research is applicable to those human images only that contain visible faces. It requires human interaction and the method that is used to estimate authenticity of an image is also prone to an estimation error.

Color discrimination has also been used as a mean to detect image forgery. To achieve that, some researchers have proposed a method called spliced image detection mechanism [9]. They detected illumination inconsistencies of an image by extracting edge or text-based features. If the image file under consideration carried information about image type, camera model, and motion after being captured, the data was found to be helpful for preventing any image forgery attempts by making the latter a difficult job [10]. However, detection of reflection-based forgeries is not a trivial task. A technique proposed by [11] suggests removing observable information from an image to make it trustworthy. Another method for detecting forgery in image files uses text-based signing of images [12]. If the digital signature gets distorted, it implies that integrity of the image had been compromised.

Thumbnails have also been used for verifying image files for authenticity [13]. The authors proposed creating thumb-

nails using contrast settings, compression, and filter models altogether which in turn are used to identify whether the actual images were compromised or not. Those models are then compared with the editing software and the originator cameras. A hidden watermark approach has also been used for image forensics [14]. It controls JPEG-lossy compression, cropping, and other possible operations that can be performed on an image by adding an invisible watermark in such a way that any distortion or a missing link in it indicates that the image had been forged.

The authors in [15] proposed an image forgery detection technique by investigating inconsistencies in lighting. Although lighting of a scene is not a complicated task, but it can be hard to match as the difference in lightings can be negligible. Researchers in [16] dealt with using a 3D lighting coefficient for image forensic. However, surface and lighting assumptions that are used are very specific. In addition to that, the challenge is to precisely estimate 3D shape of an image object.

A steganography technique to protect JPEG images from tampering by capturing two identical images instead of generating a secret text has also been discussed in [4]. The instance information is attached as a watermark to the actual image for the validation purpose. However, the proposed technique supports JPEG formats only and any slight change in camera settings between capturing images may also affect efficiency of the digital device.

Seam modification in digital images is another way of

image tampering. The former can be performed through a couple of ways; seam carving and seam insertion. In [17] the authors have studied modification of JPEG images through seam modification. A very minute change in seam effects the pixel ordering. A non-traditional method of machine learning, Classification Support Vector Machine, is used to intercept the seam-tampered image that differentiates between the tampered image and the original one. The problem with their proposed method is that it fails when highly imbalanced and skewed data sets are observed. The method is not applicable in diverse setting either.

Copy-move forgery (CMF) [18] is another common tampering technique in which a small part of the image is taken and copied to another location on the same image. Usually key-point based technology is used to detect this type of forgery, but it takes too much processing time and can run out of the memory while processing. Moreover, small cloned and smooth regions are difficult to detect. The author in [18] presents a new technique to overcome this problem. The test image is separated into smooth and rough regions and is further segmented into small regions. Before applying the Scale Invariant Features Transform (SIFT) algorithm, the customized parameters are detected for that specific image. If fixed parameters are selected to apply SIFT then results may not be satisfactory. Swarm intelligent (SI) algorithm was applied to generate a custom parameter for efficient processing of SIFT. The technique may reduce the processing time to avoid run out of memory. The experimental results indicate some higher false positive rate that needs to be improved.

In-painting [19] is another technique that has been used for forgery detection. It works by rebuilding the deteriorated part of an image. When an image gets scratched or fade away, some of its segments are reproduced to bring back its originality. The main theme was to copy segment of an image and embed it back on the scratched or deteriorated patches of the same image. The authors proposed a copy-move image forgery method in which an object is removed from an image and is pasted on a different location on the same image. Two in-painting techniques [19] were used to detect the object removal, geometry-oriented and texture-oriented. Their proposed technique, which was referred to as exemplar-based image in-painting, reported significant decrease in search time for image blocks. However, it is not very useful for multiple object removals as it increased the search overhead.

A steganography technique to protect JPEG images from tampering proposed capturing two identical images instead of generating a secret text [20]. The instance information was attached as a watermark to the actual image for validation purpose. However, their proposed technique supports JPEG formats only and any slight change in camera settings between capturing of images may also affect the efficiency of the system.

In summary, the existing approaches to counter image manipulation lack the diversity required to confront the challenge. Due to rapid rise in use of digital images, attempts to compromise their integrity are also on the rise despite currently available mitigation techniques. As it is evident from analysis of the literature, there exist no single technique that is easily applicable and equally useful to multiple types of digital images consistently; that is, computer generated images, digital

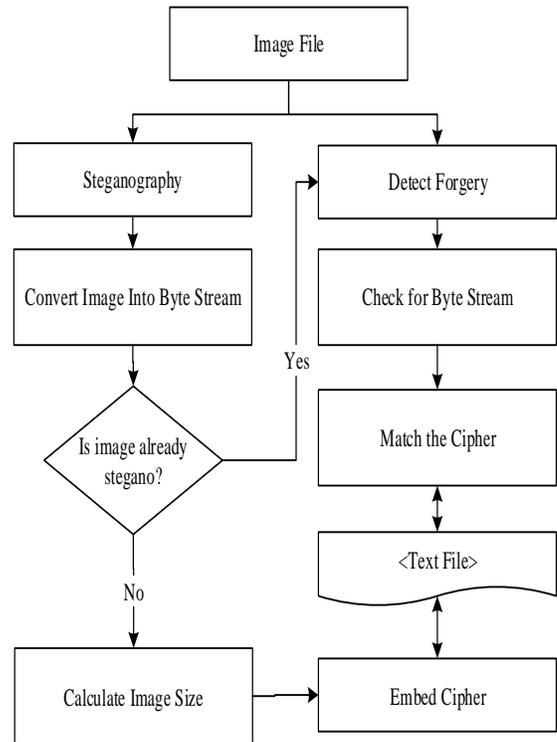


Fig. 2. An overview of Sab-iomha [6].

documents that are saved as image files, and digital camera images. The situation calls for proposing more robust methods to confront the challenge. Researchers need to come up with effective forgery detection solutions to address the issue.

III. SAB-IOMHA:THE PROPOSED TECHNIQUE

There are two phases of this research work; steganography and forgery detection. We propose a forgery detection mechanism which is a two-step approach as shown in Fig. 2. An image file is protected using an invisible watermark and then any forgeries are detected by investigating the same watermark which was inserted in the first step. As part of the approach, firstly the image is converted into byte stream that splits metadata from the file. Secondly, an invisible watermark is inserted in certain bits of the image. The watermark is in text form and can be inserted across multiple bytes. However, its length depends upon size of the image; bigger the image in size lengthier would be the watermark.

A digital image can incorporate two types of watermarks; visible watermark or invisible watermark depending upon user preferences. Visible watermark inflicts small spots on the whole image whereas the invisible one randomly inserts a text code in it. Fig. 3 is a pictorial representation of the visible watermark technique. It demonstrates different states of an image.

Visible watermarks were inserted that are noticeable by zooming the image. An ELA can identify regions within an image that possess different compression levels. It is a measure

to visually highlight difference in JPEG compression levels across different regions of an image.

Since we make use of invisible watermark, the inserted text would be hidden. We suggest composing a composite invisible watermark which is composed of multiple information fields that makes it easy to validate an image. Those fields entail cipher text, email address of the image user, and date and time stamp. At the same time the composite watermark ensures that the ownership trail of the image is maintained for any future reference as well to preserve edit history of the file. Furthermore, as part of the watermark, the cipher changes automatically if someone tries to edit the signed image as any attempts to doctor it would distort the cipher part of the inscription.

For a JPEG format, the entire image should represent the same ELA but if some fragments of an image carry different error levels, it is an indication that the original image was edited for an unauthorized modification. Regions with even coloring, like a blue or a white wall, would likely have a lower ELA levels in comparison to dark colors having high-contrast edges. For a typical forgery detection, one would check the image and try to figure out the difference between high and low contrasting edges and compare those with the ELA representation. Only a visible difference allows a naked-eye to detect any contemporary changes that might have been made to the image. Therefore, a sole ELA-dependent method is not a good fit to detect any such images which are digitally modified.



(a)



(b)

Fig. 3. Original image, and after applying a visible watermark. (b) Zoomed-in one to enhance visibility and an ELA version of the image.

In a 32-bit image that spans across four channels of colors, each pixel is constituted of four bytes. Each one of the three colors; i.e. Red, Green, and Blue is represented by a byte each as shown in Fig. 4. However, the fourth byte which is known to be reserved for Alpha does not represent anything and is

available for use. To date several systems have been proposed that represent pixels in terms of supporting colors but an ARGB is the most established arrangement for representing colors. It logically arranges a pixel in an order of Alpha, Red, Green, and Blue. As part of our composite watermark technique, we make use of the least significant bit of Alpha to steganograph an image file. This does not change data stored in any bit but text length should be calculated before it is inserted in the image file as a watermark.

Algorithm 1 Embed watermark

```

Require:  $x \geq key * 10 \vee x \neq 0$   $I = 0 || I = 10 || I = 100$ 
1:  $P \leftarrow readImagePixels$ 
2:  $P = P_0, P_1, P_2, \dots, P_n$ 
3:  $Dt \leftarrow getCurrentDateTime$ 
4:  $E \leftarrow getEmail$ 
5:  $Key = \{M_0, M_1, M_2, \dots, M\}$ 
6:  $x \leftarrow key + Dt + n$ 
7:  $x \leftarrow floor(\frac{x}{k(k+1)})$  equation 1
8: function MATCHCIPHER( $key, P$ )
9:   if found then
10:     return
11:   end if
12: end function
13: function INSERTCIPHER( $x, P$ )
14:   function INSERTEMAILANDDATETIME( $Dt, E, P$ )
15:     for  $j = 1$  to  $j = 8$  do
16:       StegPixel
17:     end for
18:   end function
19:   for  $i = Dt + E$  to  $i = x$  do
20:     for  $j = 1$  to  $j = 8$  do
21:       StegPixel
22:     end for
23:      $x \leftarrow x + I$ 
24:   end for
25: end function

```

Fig. 4 demonstrates how exactly our proposed technique makes use of certain bits of an image file. It splits metadata from the file header. The file is then converted into pixels which in turn is transformed into a byte stream. Alpha bits are selected, and an invisible watermark is inserted into them, which is a composition of cipher, email address, and time and date stamp. If we consider an image as a matrix P having m rows and n columns, total number of pixels in it can be determined using the given m n relation.

We argue that inserting watermark into the least significant bit is an easy yet effective approach for signing an image with the traceable information. Eighth bit of the Alpha bytes is utilized for the purpose; i.e. one bit of the overall size of the inserted watermark. It should be noted that we do not make use of all Alpha bytes of an image file. Their selection is based on a certain pattern which is generated at run time to ensure maximum protection of the image. For a four-byte image having thirty-two bits, the least significant bit of the Alpha component is utilized which is depicted as the marked bit of a pixel shown in Fig. 4(d). An image consisting of 800 600 pixels can store up to 1,440,000 bits or 180,000 bytes of watermark. For instance, a block of 8 pixels of a 4-byte image can be represented as: if number 35 is inserted as a watermark

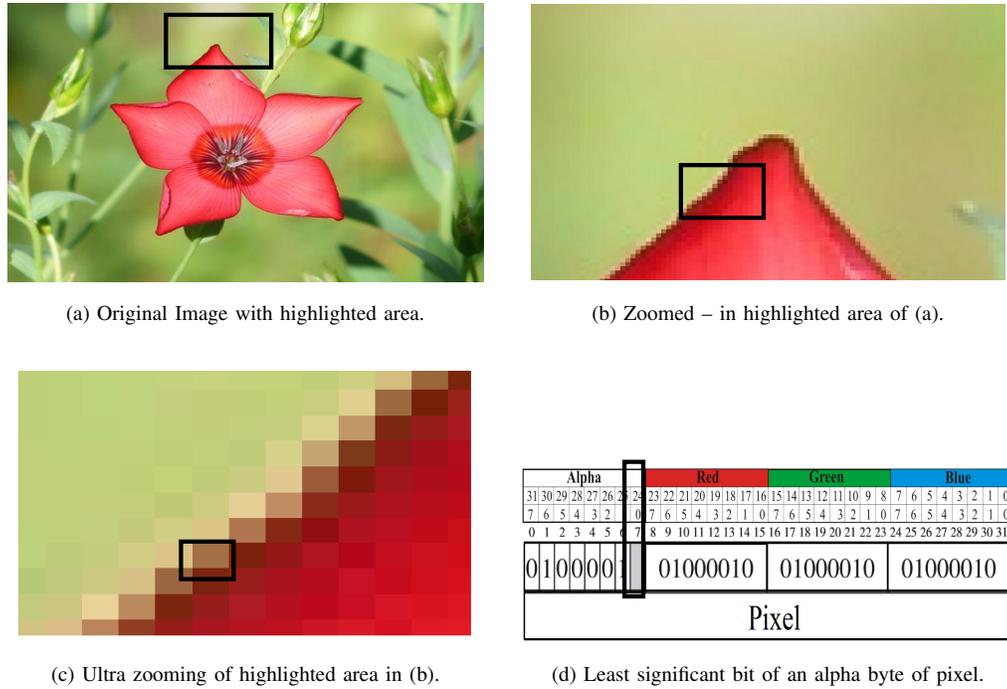


Fig. 4. Illustration of an image pixel and the corresponding bit used for the invisible watermarking.

having binary representation 00100011 across Alpha bits of an image, the resulting pixel block gets manipulated in such a way that 35 is accommodated in consecutive pixels highlighted as shaded pixel bits in Fig. 5. It is worth mentioning that only least significant bits of Alpha bytes are inserted with the watermark fragments. All pixels can be protected using the scheme which does not affect the visual contents of the image file. Since the proposed technique consumes an image at the structural level, its steganography cannot be observed through a naked eye.

Email	Date Time	Intensity	Cipher length
6-255	7	1	n * key ...

$$(y - 1)k + (a - 1)k^2 < N < (y - 1)k + ak^2 - 1 \quad (1)$$

In a 32-bit colour image, Alpha bits are separated, and the code stream is spread across the byte stream using Algorithm 1. Where x is the number of pixels in an image, I is intensity of the watermark which can be 10, 50 or 100, and Key is length of the cipher. P is an array of pixels which an image file contains.

Dt is the current date and time of the system. E is email address of the user. At line 7 of the algorithm, x is cumulation of the composite watermark obtained by adding cipher text, date and time stamp, and email address. The cipher text constitutes the constant part of the watermark whereas rest is the system and user dependent to enhance the strength of the algorithm. The function at line 8 checks the image file for the watermark, if matched, the image is authenticated. Otherwise, InsertCipher procedure at line 13 is initiated. The cache space

can be increased to any positive numeric value in case we want to add an interval between the bytes that are occupied by the ark.

There could possibly be a case that someone else signs the image after it was steganographed by the actual author. The situation makes it nontrivial to keep track of the actual ownership. The combination of date and time in particular ensures that once a user signs the image, the ownership trail can be maintained for the subsequent detection of any successful forgery attempts. Table I illustrates the composition of the composite watermark. Email address of the user is allocated up to 255 bytes, date and time is allocated 7 bytes, 1 byte for Intensity which is the distance between two nearest cipher bytes, and variable number of bytes are reserved for the Key which points to the cipher text. The following equation I is used for determining the length of the cipher.

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} & \dots & P_{1n} \\ P_{21} & P_{22} & P_{23} & \dots & P_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ P_{m1} & P_{m2} & P_{m3} & \dots & P_{mn} \end{bmatrix} = (p_{ij})_{m \times n}$$

Where a is any positive integer and y is the cumulative length of characters of email address and date and time stamp, K is constant length space allocated for the cipher text to be impeded in the image, and N represents length of the image in bytes.

IV. AUTOMATION OF SAB - IOMHA

The software tool that we developed to automate our research is relatively simple and user friendly with minimum

0	1	0	0	0	0	1	0	01000010	01000010	01000010	0	1	0	0	0	0	1	0																		
0100001							01000010							01000010							01000010															
01000011							01000010							01000010							01000010															
0	1	0	0	0	0	1	0	01000010	01000010	01000010	0	1	0	0	0	0	1	0																		
01000010							01000010							01000010							01000010															
01000010							01000010							01000010							01000010															
.....skip x bytes Algo. 1 line 23																							01000010							01000010						
01000011							0100001							01000010							01000010															
Replacement of least significant Alpha bits with cipher bits.																																				

Fig. 5. Least significant bits of Alpha bytes of an image.

of work-flows. It supports browsing of an image file using a GUI interface and is loaded in computer memory.

Fig. 6 depicts user interface of the tool we developed. It was programmed using Java technologies. The ultimate objective is to facilitate validation of digital images and documents as well in case they are in an image format to prove integrity of the contents or to verify that the digital document has not been edited since its creation. The tool supports multiple features as shown in Fig. 6. The Steg Image embeds an invisible watermark in the image. The steganographed image can also be saved on the disk for any future reference. Forgery Detection opens up another screen as depicted in Fig. 6.

Signing an image file is a two step procedure: in the first phase, we would steganograph an image by inserting the invisible watermark which is validated for integrity in the second phase. We randomly pick an image and upload it to the tool to demonstrate usefulness of our technique as well as the overall automation itself. The sample image on the right side of the Fig. 7 is signed using the watermark which is the composition of cipher text, email address, and date and time stamp. It can be observed that quality of the image was not compromised at all by using the technique. The same file can be checked to verify if the image is original or any attempts has been made to alter it. In case the validation procedure generates an alert text, which is the case as shown in Fig. 7, it is an indication that the image has been forged by some other user. Otherwise, the inserted watermark is displayed to testify the originality of the image. Algorithm 2 enlists steps performed to detect forgery. It is a three-step procedure; in the first one, it looks for an insertion, if not found, it implies that the image is not steganographed. If an insertion is found, it is matched with the actual watermark. If the exact match is not found, the image is reported to be forged. Otherwise, it is the original one.

To further validate the proposed technique, we performed an experiment to demonstrate its effectiveness. A set of images with varying range of size was steganographed using the tool we have developed to automate Sab-iomha. The motivation was to compare metadata of the image files before and after the technique was applied. We considered certain factors like size, compression level, and resolution to investigate the subject. Each image had 4 color channels having 32 bits altogether,

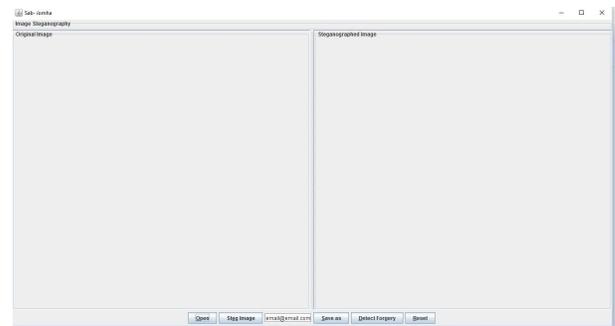
Algorithm 2 Forgery Detection

```

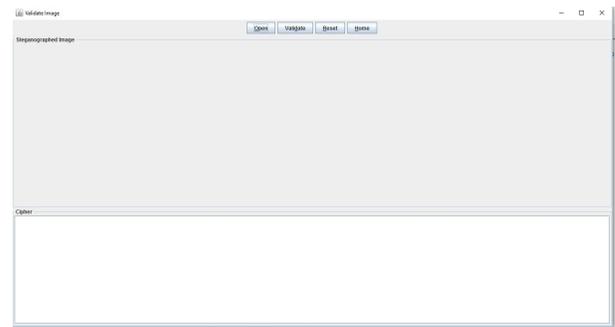
Require: key , image
1: P ← readImagePixels
2: P = P0, P1, P2, ...Pn
3: Key = {M0, M1, M2, ...M}
4: function MATCHCIPHER(key, P)
5:   if found then
6:     if key = extractedCipher then
7:       Image is original
8:     end if
9:   if key ≠ extractedCipher then
10:    Image is forged
11:  end if
12:  Key = ImageCipher
13:  Original Image
14:  return
15: else if
16:   thenImage is not protected
17: end if
18: end function

```

and 0.27 value for mega pixels. Table I reflects upon the image population in more detail.



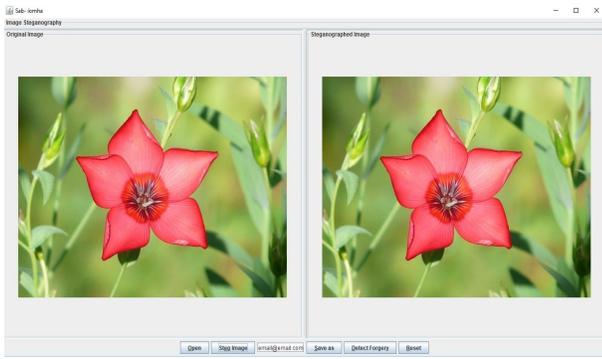
(a)



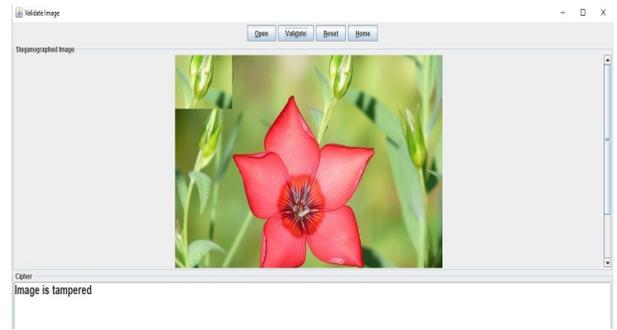
(b)

Fig. 6. (a)Home-interface of the tool implementing Sab - iomha. (b) To detect an image le for forgery [6].

Table I draws comparison between metadata of the image files before and after applying the steganography using Sab-iomha. It is noticeable that color type remained the same even after each image was steganographed, that is, RGB with Alpha. There was no change in resolution of the images either. However, some difference was observed in terms of size of each image. In general, the steganographed images were noted



(a)



(b)

Fig. 7. Home-interface of the tool with an image loaded and steganographed.

TABLE I. POPULATION OF THE IMAGE FILES FOR EXPERIMENTATION

No.	Original Image		Processed Image	
	Size (Kilo Bytes)	Resolution	Size (Kilo Bytes)	Resolution
1	655	600x450	648	600x450
2	291	457x360	435	457x360
3	511	600x450	502	600x450
4	914	1280x1012	1152	1280x1012
5	129	262x192	129	262x192
6	317	425x281	313	425x281
7	1238	1024x768	1168	1024x768
8	89	284x177	90	287x177
9	726	700x350	725	700x350
10	1525	1024x750	1492	1024x750
11	393	476x500	366	476x500
12	136	276x183	135	276x183
13	590	600x450	581	600x450
14	364	500x334	358	500x334
15	158	259x194	159	259x194
16	139	259x194	139	259x194
17	129	275x183	128	275x183

to be slightly smaller in size. The overall analysis suggested that quality of each set of images remained the same, i.e. studying the metadata before and after the application of the forgery detection technique did not negatively influence the quality of the images under consideration.

V. CONCLUSION

Digital images are prone to forgery in the current age as it has become much easier to manipulate digital contents due to advancement in the domain. We have introduced a new dimension to the digital image steganography by proposing a light weight technique. It uses a composite watermark to check digital images for authenticity. The proposed technique signs digital images for integrity and protects them against any manipulations. The forgery issue is addressed in a novel way; ELA, JPEG, and metadata are incorporated, and an invisible watermark is inserted to enhance efficiency and effectiveness of forgery detection. The proposed technique is automated through a software tool which facilitates users to steganograph digital images. The same image can then be checked for originality. The core purpose of the tool development is to support the usability of Sab-iomha which may not only validate

photographs but also any digital contents stored in an image format. This work enables even non-technical users to be able to investigate integrity of image files at their own. It also empowers them to get insight on their digital contents. As part of the validation mechanism, we have tested the algorithm on a series of random images. The results suggested that the technique can not only verify the digital images for authenticity but also does not negatively influence their quality. Moreover, users can also protect their images from any attempts to forge. The research we conducted do not have any ethical, moral and legal issues associated with it. The project is economically feasible too as the users do not require to purchase any hardware devices and are alleviated from the need for software installations. Currently, the work is aimed at supporting JPEG and PNG file formats only. We aim to extend support for other image formats in the future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their feedback which helped to improve the earlier version of the research paper.

REFERENCES

- [1] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.
- [2] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417.
- [3] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho, "Humans are easily fooled by digital images," *CoRR*, vol. abs/1509.05301, 2015. [Online]. Available: <http://arxiv.org/abs/1509.05301>
- [4] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, Dec 1997.
- [5] T. J. d. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. d. R. Rocha, "Exposing digital image forgeries by illumination color classification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 7, pp. 1182–1194, July 2013.
- [6] M. S. Bhatti, S. A. Hussain, A. Qayyum, I. Latif, M. Hasnain, and S. I. Hashmi, "Sab - iomha: An automated image forgery detection technique using alpha channel steganography," in *Recent Advances in Information Systems and Technologies*, A. Rocha, A. M. Correia, H. Adeli, L. P. Reis, and S. Costanzo, Eds. Cham: Springer International Publishing, 2017, pp. 736–744.

- [7] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, March 2009.
- [8] W. Fan, K. Wang, F. Cayre, and Z. Xiong, "A variational approach to jpeg anti-forensics," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 3058–3062.
- [9] H. F. Matthias Kirchner, Peter Winkler, "Impeding forgers at photo inception," pp. 8665 – 8665 – 9, 2013. [Online]. Available: <https://doi.org/10.1117/12.2008412>
- [10] J. F. O'Brien and H. Farid, "Exposing photo manipulation with inconsistent reflections," *ACM Trans. Graph.*, vol. 31, no. 1, pp. 4:1–4:11, Feb. 2012. [Online]. Available: <http://doi.acm.org/10.1145/2077341.2077345>
- [11] V. Conotter, G. Boato, and H. Farid, "Detecting photo manipulation on signs and billboards," in *2010 IEEE International Conference on Image Processing*, Sept 2010, pp. 1741–1744.
- [12] E. Kee and H. Farid, "Digital image authentication from thumbnails," in *Media Forensics and Security II*, vol. 7541. International Society for Optics and Photonics, 2010, p. 75410E.
- [13] C.-T. Hsu and J.-L. Wu, "Hidden digital watermarks in images," *IEEE Transactions on Image Processing*, vol. 8, no. 1, pp. 58–68, Jan 1999.
- [14] W. Luo, Z. Qu, F. Pan, and J. Huang, "A survey of passive technology for digital image forensics," *Frontiers of Computer Science in China*, vol. 1, no. 2, pp. 166–179, May 2007. [Online]. Available: <https://doi.org/10.1007/s11704-007-0017-0>
- [15] M. K. Johnson and H. Farid, "Exposing digital forgeries in complex lighting environments," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 450–461, Sept 2007.
- [16] W. Fan, K. Wang, F. Cayre, and Z. Xiong, "3d lighting-based image forgery detection using shape-from-shading," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Aug 2012, pp. 1777–1781.
- [17] K. Wattanachote, T. K. Shih, W. Chang, and H. Chang, "Tamper detection of jpeg image due to seam modifications," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2477–2491, Dec 2015.
- [18] F. Zhao, W. Shi, B. Qin, and B. Liang, "Image forgery detection using segmentation and swarm intelligent algorithm," *Wuhan University Journal of Natural Sciences*, vol. 22, no. 2, pp. 141–148, Apr 2017. [Online]. Available: <https://doi.org/10.1007/s11859-017-1227-4>
- [19] Z. Liang, G. Yang, X. Ding, and L. Li, "An efficient forgery detection algorithm for object removal by exemplar-based image inpainting," *J. Vis. Commun. Image Represent.*, vol. 30, no. C, pp. 75–85, Jul. 2015. [Online]. Available: <http://dx.doi.org/10.1016/j.jvcir.2015.03.004>
- [20] T. Denemark and J. Fridrich, "Steganography with multiple jpeg images of the same scene," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 10, pp. 2308–2319, Oct 2017.

Recommendations for Building Adaptive Cognition-based E-Learning

Mostafa Saleh, Reda Mohamed Salama
Information Systems Department
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Abstract—Adapted e-Learning systems try to adapt the learning material based on the student’s preferences. Course authors design their courses with their students’ styles and in mind, course delivery should match the student style, and student assessment should also be adapted to match each specific student’s learning style, while student portfolio helps identifying the student model. To the best of our knowledge, no clear recommendation for building community wide adapted and personalized e-learning systems. This paper presents recommendations to add adaptation and personalization to one of the most common open source Learning Management System (LMS), Moodle. The adaptation features are based on using learning styles, ontology, and cognitive Bloom Taxonomy in building and presentation of the e-learning material (Learning Objects). This is helpful to establish adaptable and cognition-based Learning Object repository and course development centers.

Keywords—Adaptive e-learning; learning objects, learning styles; student models; open source LMS; Moodle; personalized teaching model

I. INTRODUCTION

E-Learning is taking a great attention worldwide. It is supposed to contribute to enhance the traditional education if properly implemented. It can be beneficial to most forms of e-Learning, e.g., training, girls’ education, continuing education, open education. It can even be used as a supporter and enhancer for traditional in-class education.

As each learner has different learner’s characteristics; so, utilizing diverse educational settings may be more appropriate for one group of learner than for another. So, adaptive e-learning is an e-learning system that is more effective by adapting or personalizing the presentation of information to individual learners based on their preferences, knowledge and needs. This sort of e-Learning systems tries to acquire knowledge about a particular learner and offer personalized services and enable one-to-one delivery [1], [2].

Learners are the main actor in the e-Learning environment and they are usually having varied and diverse cognitive and psychological traits. One of the important facets of the adaptive model of e-Learning is to adapt the presentations of the learning material to meet the needs of each individual learner during the course delivery process. To achieve such goal, we need to detect the learner profile to adapt the content and presentation of the learning material. This profile is called

Student Model (SM). Also, the learning materials are composed of small granular multimedia objects referred to as Learning Objects (LOs), to achieve a high level of adaptation.

Student model should be used for tailoring the teaching strategy and learning material for dynamically adapting it according to the student’s abilities and his/her previous knowledge. Student model is often based on various different dimensions. In this project, we focus on the student model in one dimension, namely, the cognitive model, especially the learning style. A learning style is defined, among many definitions, as “the unique collection of individual skills and preferences that affect how a student perceives, gathers, and processes learning materials” [3]. Therefore, the concept of student model, especially learning styles, is considered as a central component in this research’s implementation. Course authors should design their courses with their students’ styles in mind, course delivery should match the student style, and student assessment should also be adapted to match each specific student’s learning style, while student portfolio helps identifying the student model.

Learning Objects are stored in what is called Learning Objects Repositories (LOR). Learning objects are drawn from an LOR based on a certain criterion, which is described in terms of metadata attributes that are used to specify the selection criteria of the appropriate required material. In this research we suggested adapting the LO metadata of a standard LO model such as SCORM by adding extra attributes necessary for supporting the concepts of the student model, especially the dimension of the learning styles.

Learning styles mean that individuals differ in regard to what mode of instruction or study is the most effective for them [4]. So, they are distinct individual patterns of learning that vary from person to person. It is necessary to determine what is most likely to trigger each learner’s concentration, how to maintain it, and how to respond to his or her natural processing style to produce long term memory and retention [5].

There are many learning style models exist in literature, e.g. the learning style model by Felder and Silverman [6], Kolb [7], Mumford and Honey [8]. They agree that learners have different ways in which they prefer to learn. After a comprehensive study of the e-learning environment, we selected Felder and Soloman’s Index of Learning Styles (ILS) [9].

Bloom's Taxonomy of the Cognitive Domain Bloom's taxonomy is possibly one of the best known and most widely used models of human cognitive processes [10]. It includes Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation levels. A revised version of the taxonomy was published in 2001 [11].

The adaptive and artificial intelligent tutoring systems (ITS) are developed using Web 2 [12]. The systems are developed to adjust the contents as per the effective learning styles that are identified using self-organizing maps (SOMs). Artificial intelligent systems behave like human beings. Supervised, unsupervised and reinforced are three types of artificial intelligent systems. Supervised system needs examples and a teacher to train. Unsupervised system is trained without a teacher and it rectifies itself after a mistake is reported. Reinforced system needs a mentor to guide the system that the answers are true or not. Unsupervised learning is selected to train the tutoring systems because it does not require a teacher and Felder-Silverman Learning Style Model (FSLSM) are used. The intelligent and adaptive tutoring systems are equally portable to run on web and mobile platforms.

Intelligent educational system (INES) is one of the components of an electronic learning platform [13]. Semantic management of users and contents, BDI-based (believes, desires, intentions) agent, an inference engine, ontologies and learning contents are the main components of INES. INES is used to identify the credentials of each student and check the status of his/her learning progress. The core objective of this exercise is to recommend a student whose progress of learning is not satisfactory.

An intelligent information access system (IIAS) is engineered to introduce new learning theories for the undergraduate students [14]. Concept, case and internet based learning theories are taken into account while developing the proposed system. IIAS identifies and marks important notes about an experimental medical case and it also assists in conducting objective assessments. The complexity of test assessment can be tailored according to the semester number of a student.

A similar study is conducted while developing an educational system [15]. The proposed system depends on abilities of students and degree of interaction between students and instructor [15]. The proposed educational system uses multi agent domain ontology to measure the progress of learning and judge the abilities of a student. A student interacts with the system to describe his/her opinion about a topic and it is matched with the data of text book. The system displays the mistakes of a student and it also suggests improvement in the course material.

Escudero and Fuentes [16] propose a general purpose tool that can be used to design courses for an intelligent tutoring system regardless of the platform. The platform independent courses are interoperable to easily call and use. The idea of such courses will help the practitioners to deal with a single course as an independent software component and it will incorporate known advantages of component based

development into ITS such as reusability, time saving and economical.

A study is conducted to propose a new idea to deal with the (ITS) [17]. Escudero and Fuentes [17] propose a general purpose tool that can be used to design courses for an intelligent tutoring system regardless of the platform. The platform independent courses are interoperable to easily call and use. The idea of such courses will help the practitioners to deal with a single course as an independent software component and it will incorporate known advantages of component based development into ITS such as reusability, time saving and economical. The general purpose tool is tested using two ITSs to conclude the results.

El-Bishouty et al. [18] recommend developing an e-learning system that is intelligent, adaptive, and customizable. The proposed e-learning system should have the features to generate courses and recommend improvements according to the level of interacting student. The proposed system will use behavior, learning style and cognitive skills (BLC) of a student to train. El-Bishouty et al. [18] recommend that it is vital to consider BLC as a basic building block to develop an effective and efficient e-learning system to achieve the desired results.

A research is conducted to model and adapt the user in a virtual environment [19]. CUMULATE is a general purpose student modeling server that is developed by Brusilovsky et al. [19] to describe the e-learning architecture and knowledgetree in a distributed environment. Knowledgetree is software that is used to provide online services. Subject based search is used to infer using CUMULATE and QuizGuide during the self-evaluating quizzes.

By investigating these systems, we can conclude that: None of the above literature addressed the following subjects which shape the objectives of our research:

- 1) Automatic generation of a course syllabus, Table of Contents (TOC), and course material.
- 2) Automatic adaptation of the course syllabus: generation of adapted course syllabus, adapted TOC, and adapted course material according to the student background knowledge.
- 3) Adapted course delivery according to the student model.
- 4) Adapted student assessment: placement of quizzes during the course, assessment of prerequisite knowledge, post course assessment according to the student model, especially the student learning style.
- 5) Integrating the concept of Bloom's taxonomy to enhance the comprehensiveness of the domain ontology. This adoption and enhancement to domain Ontology affects all the learning components of authoring, delivery and assessment.
- 6) Utilizing m-Learning to the system.
- 7) Support tools for building LORs and creating LOs from existing learning material.
- 8) Adaptive open source LMS.

In this paper, we present recommendations to add adaptation and personalization to one of the most common

open source Learning Management System (LMS), Moodle. The adaptation features are based on using learning styles, ontology, and cognitive Bloom Taxonomy in building and presentation of the e-learning material (Learning Objects). This is helpful to establish a nation-wide adaptable and cognition-based Learning Object repository and course development centers. The rest of the paper is organized as follows: Section 2 presents the adaptive e-Learning System (KAU-AES) developed at King Abdulaziz University. Section 3 is directed to the knowledge base building recommendations and Section 4 presents the recommendations of the authoring system. Section 5 discusses the adaptive course delivery system recommendation. Section 6 gives the recommendations for assessment system, and finally Section 7 presents the discussion and conclusion.

II. ADAPTIVE E-LEARNING SYSTEM: KAU-AES

The major objective of this paper is to give recommendations based on theoretical and practical experience to build adaptive e-Learning environment community. Instead of building an environment from scratch to support all the educational services required by the educational institutions, we used Moodle because of its popularity as it is used in several universities. Moodle also is known as simple and easy to adapt and customize to the needs of the educational system. Therefore, Moodle is integrated to many of the components that were developed to compose the Adaptive e-Learning Environment, as shown in Fig. 1.

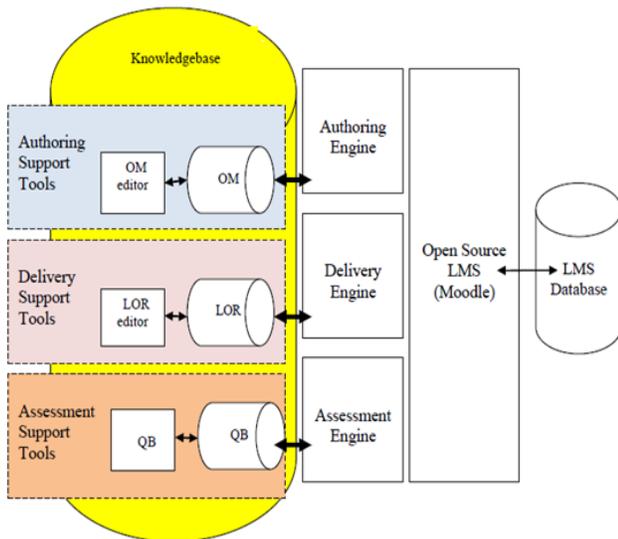


Fig. 1. High Level Architecture of the Adaptive e-Learning Environment

Three main subsystems in the proposed adaptive e-Learning system are integrated to the open source Moodle, namely, Authoring, Delivery, and Assessment engines. Each of those main engines works smartly with the aid of the Knowledge base. This knowledge base, in turn, is composed of three main knowledge bases, namely, the Ontology Model (OM), the Learning Object Repository (LOR), and the Question Bank (QB), each of which is maintained with the aid of a specially designed editor. Finally, the normal database of Moodle is updated to accommodate more data as required by the adaptive environment, such as:

The student information is updated to accommodate the Student Model (SM) by adding both his/her background knowledge, learning style model, and some other data, such as, preferred language, etc. The course information is also updated to include the Course Learning Outcomes (CLO). In addition, the Moodle itself is adapted to accommodate and seamlessly integrate to the different components of the adaptive e-Learning Model. For instance, the following was implemented to augment Moodle with adaptation:

The student page is updated to allow for editing and updating the student model.

The Teacher page is adapted to allow him to edit the course LOs, CLOs, and Generate the Course Syllabus.

When the student registers in a course, the course CLOs are automatically adapted to suit this specific student according to his/her student model. His/her course syllabus and course table of contents are adapted accordingly. Therefore, the Moodle page for the student is adapted to display the student adapted CLO, the adapted Course Syllabus, and the detailed adapted Course TOC.

The following steps are recommended to build the adaptive e-Learning Environment: Design and build the core Knowledge base.

Design and build the knowledge base and tools:

- The LO specification and meta-data structure.
- The Domain Ontology network structure so as to augment Bloom's Taxonomy.
- The Student model components and dimensions.
- Design and build the authoring support tools.
- Design and build the adaptive delivery engine.
- Design and build the assessment engine.

III. KNOWLWDGE BASE BUILDING RECOMMENDATION

All components of the adaptive e-Learning Environment are centered on the knowledge base. As shown in Fig. 2, the Knowledge base is composed mainly of three major components: the system knowledge base; student database; and course database. The Knowledge base is composed of the Learning Object Repository (LOR) and the Ontology Model (OM). While, database is composed of the Student Model (SM) and the Course Model (CM), which themselves are further decomposed. The SM is composed of two components: the student's Learning Style Model (LSM) that is defined in terms of the four dimensions of FSLSM [6] and the SBDK representing the knowledge that the student captures with an acceptable cognitive depth for the domain of study. In addition, the CM is composed of three components: the CLO, the Course Syllabus, and the TOC.

Two database components that are essential to the adaptive processes, namely, the Student Model and the Course Model, which maintains data along those two models for each student and each course, respectively. The student model has two major components in addition to few other attributes. The

course model has three components; each is having two levels, generic and adapted to suit each student.

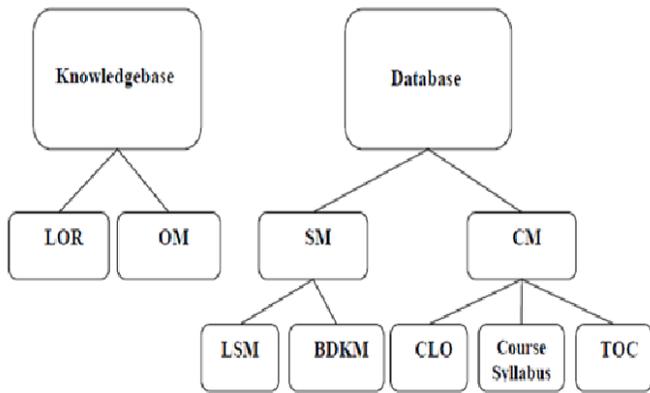


Fig. 2. The Knowledge base

The Student’s Learning Style Model (LSM): Each student has his/her own learning style model which is defined in terms of the FLSM’s four dimensions (Visual/Verbal, Global/Sequential, Active/Reflective, Sensing/Intuitive). The LSM is identified for each student once, at the time he/she joined the e-Learning system. The LMS is identified through the index of FLSM questionnaire (<http://www.engr.ncsu.edu/learningstyles/ilsweb.html>) which is considered an easy way to identify the learner’s learning style in more details. This questionnaire contains 44 questions and describes the learning style dimensions by using scales from -11 to +11; while zero indicates the origin of the axis, each direction on the axis refers to one of the two properties of the dimension.

Instead of asking the student to fill the questionnaire in sequence (the 44 questions), we grouped the questions related to each two dimensions in a single group of questions as shown in Table I. From the practical experience with students while they are filling the questionnaire, this enables them to choose the most related to their preferred learning style as they sometimes find some confusion in understanding each question separately.

The Course Model is composed of three components, two of which, namely, the course syllabus and the Course Learning Outcomes that are defined by the course designer. Moreover, the Course Model has two levels of data: the highest level is more generic and concerns the course from a generic perspective, i.e., one course fits all, while the other is the adapted course for each individual student according to his/her Student Model.

This generic course model is simply a course syllabus that is automatically generated from the course’s CLOs with the aid of the Domain Knowledge Ontology Model. It is generated for all students with no guarantee it matches the student model of any of the students. In addition, the course’s generic TOC is automatically generated to match the teacher’s teaching style. On the other hand, the lower level of data of the Course Model are the adapted Student’s CLO, the adapted Student Course Syllabus, and the adapted Course TOC, which are

adapted for each individual student according to his/her student model.

TABLE I. GROUP SELECTION FOR FELDER LEARNING STYLE DIMENSIONS (ACTIVE/ REFLECTIVE/NEUTRAL)

A	B	C
Active	Reflective	Neutral
<p>I understand something better after I try it out.</p> <p>When I am learning something new, it helps me to talk about it.</p> <p>In a study group working on difficult material, I am more likely to jump in and contribute ideas.</p> <p>In classes I have taken I have usually gotten to know many of the students.</p> <p>When I start a homework problem, I am more likely to start working on the solution immediately.</p> <p>I prefer to study in a study group.</p> <p>I would rather first try things out.</p> <p>I more easily remember something I have done.</p> <p>When I have to work on a group project, I first want to have "group brainstorming" where everyone contributes ideas.</p> <p>I am more likely to be considered outgoing.</p> <p>The idea of doing homework in groups, with one grade for the entire group, appeals to me.</p>	<p>I understand something better after I think it through.</p> <p>When I am learning something new, it helps me to think about it.</p> <p>In a study group working on difficult material, I am more likely to sit back and listen.</p> <p>In classes I have taken I have rarely gotten to know many of the students.</p> <p>When I start a homework problem, I am more likely to try to fully understand the problem first.</p> <p>I prefer to study alone.</p> <p>I would rather first think about how I'm going to do it.</p> <p>I more easily remember something I have thought a lot about.</p> <p>When I have to work on a group project, I first want to brainstorm individually and then come together as a group to compare</p> <p>I am more likely to be considered reserved</p> <p>The idea of doing homework in groups, with one grade for the entire group, does not appeal to me.</p>	

The student’s BDKM is used to adapt the student’s CLO and Course Syllabus, while his/her LSM is used for adapting the Course TOC. The Course CLO represents the goal outcomes of this course as specified by the course designer. It takes the form of a list of items, each of which is described as follows:

“By the end of this course the student will be able to: <Revised Bloom Taxonomy cognitive level> the <Concept name/id> at a complexity level of <depth level>”.

For example,

“By the end of this course the student should be able to Apply the concept of Stack at a complexity level of 2”.

With the aid of the OM, the generic course syllabus is generated. The syllabus is composed of numbered sections which in turn are composed of subsections, while the TOC adds sub-subsections which go into pedagogical details. For instance, a section on Stack may contain a subsection that explains the concept of “LIFO”, while the TOC may further break down the “LIFO” subsection into many sub-subsections, like an definition, an application of LIFO from real life, etc.

Using the Background Domain Knowledge Model (BDKM) of the Student Model of a certain student, the CLO will be adapted to match this specific student (hence is named Student Learning Outcomes (SLO)) by adding unknown prerequisite concepts, and removing well known concepts. Again, the Authoring System will use the adapted SLO, with the aid of the OM, to automatically generate the adapted course syllabus, which will then be the input for generating the adapted student's course TOC.

The Cognition-Augmented Knowledge base has two main components, namely, the LOR and the Ontology Model (OM). Those two components are main drivers of the adaptation. OM derives the Authoring Process, while LOR derives the Delivery Process. Both components play an important role during the pre and post assessment processes.

A subject matter expert course author, who is very much familiar and knowledgeable about the subject domain knowledge, knows much invaluable information about those concepts and the best ways of teaching them to a certain group of students with a specific average profile. For instance, the expert author should know what the best break down is for a certain specific topic; what the best sequence for certain topics would be; what topics would achieve the goals of a certain course; what the best depth is for each topic/subtopic; when to introduce exercises, quizzes, and tests, etc. to stimulate students' enthusiasm and learning effectiveness. One goal of this research is to support course authors in doing the authoring job professionally, even if they lack the sufficient expertise.

In the e-Learning Model, that in-depth knowledge regarding a specific knowledge domain is accumulated in the OM, which is assumed to be incrementally and/or cooperatively designed by the domain experts. In fact, OM is a Key Player in the e-Learning Model. It is a comprehensive model of interrelationships among concepts/topics. This comprehension gives more flexibility to the authoring process in composing a course. Moreover, it gives an automation power to the authoring process.

So, it is recommended to design OM with the objective of supporting not only course authoring but also course delivery and assessment as well. To achieve this goal, the traditional Ontology net scheme is extended to accommodate two extra updates to the classical scheme:

- Adding a measure of depth/complexity to each concept node in OM [20].
- Embedding the concepts of the instructional design theories and the Revised Bloom's Taxonomy [21, 22].

Complexity Level Extension: In OM, the concept's node is a complex structure. Each node is given a complexity value (F=Fundamental | M=Medium | D=Advanced) that is intended to guide the design of a course according to the course's complexity. To explain, a 200-level course wouldn't have the same topics/concepts as those higher-level courses; as the course level increases as the complexity of the concepts increases.

However, usually a higher-level course would also introduce those concepts of a lower complexity. Therefore, for the navigation through the OM net during the course design processes, it is recommended to use the following simple rule.

In a course of a complexity level "c", all concepts of a complexity higher than "c" wouldn't be included in this course. For instance, if the course is a medium-level course, all advanced concepts (Marked with D) would be ignored; only F & M concepts are included.

Embedding RBT in OM: The second improvement in OM is the accommodation of the RBT [22]. Each concept node is made of six levels corresponding to Bloom's levels. This will make OM as a multilayered diagram; one layer for each of the Bloom's levels. This extension is intended to guide the course design phase in which the course objectives specify the target Bloom's level for each concept covered in the course. Accordingly, this concept's OM's layer is employed and the relationship links are followed. Most importantly of those links is the prerequisite link which might reference a specific layer of another concept, as shown in Fig. 3, where the "Depth-Limiting Search strategy", for instance, is having complexity level "M" and whose RBT's level of "Understanding" requires, as a prerequisite, "Depth-First Search" at RBT's Level of "Applying".

Noteworthy, not only the course authoring is intelligently impacted by the extended OM but also many other components in the Knowledge base. For instance, the student's BDKM is updated to accommodate the six levels of RBT. Accordingly, OM plays an important role in the adaptation of the course delivery in two ways:

- A more accurate evaluation of the student knowledge as compared to the prerequisite requirements, and
- Compensation of missing prerequisite knowledge.

This feature is implemented with only the first tree levels in the RBT and the compensation of the missing prerequisite knowledge is done through "recall" branch as in Fig. 3.



Fig. 3. Learning Object Folder Structure

Each Learning Object is described, and hence selected, using a set of metadata attributes. The LO Metadata Model extends the standard metadata model of SCORM by adding few extra attributes to accommodate the adaptation theme of the e-Learning Model. In other words, the LO Model has extended the standard metadata model of SCORM by:

1) Adding extra attributes necessary for supporting the theories it implements, such as Learning Style Model, Revised Bloom's Taxonomy, etc. Of course, these attributes are not contradicting with any LO standard, but rather they are complementing them,

2) Employing some of the SCORM's attributes after stretching their space of acceptable values.

In general, these Metadata Attributes are used for two main purposes:

3) Searching and retrieving the LOs easily and precisely either manually or automatically.

4) Aiding in the process of adaptation and personalization through choosing the proper LOs meeting specific criteria.

The Metadata Model: The adaptation process applies different theories such as Learning Style, instructional design, and cognition theories, a knowledge that are usually applied by an expert instructor who happened to know them through study or by experience. Inexpert instructors, on the other hand, though are subject matter experts, usually lack such knowledge. The e-Learning Model attaches a set of metadata attributes to each LO in order to aid the adaptation process. Those attributes are so simple and naive in such a way that they don't require an expert to define them, yet are used by the expert system to deliver courses with a similar quality like that of an expert instructor. Each LO is described in terms of several metadata attributes.

IV. AUTHORING SYSTEM RECOMMENDATION

Once the course is added to the system and its CLOs are defined, the algorithm of the "Generic Course Syllabus Generator" runs to generate the generic course syllabus, while the algorithm of the "Adapted Student Course Syllabus Generator" runs once the student registers in a specific course. The files are placed in an agreed upon folder and named with an agreed upon naming convention. The idea of the Generic Course Syllabus Generator is summarized as follows:

1) For each Concept in the CLO, consult OM to identify its "ISA" and "Prereq" linked concepts. Those concepts should be added to the syllabus before the concept as "recall" concepts so that they are briefed to the student before start teaching the concept itself.

2) For each Concept, its "ComposedOf" (or sometimes called "PartsOf") relationships in OM are followed to identify the breakdown of this concept.

3) Those subordinates of the concept (its ComposedOf concepts) are ordered using the "follow" relationships among them.

4) Only subordinate concepts of complexity level less than or equal the level specified in the CLO are considered.

5) This procedure is recursively done so that the same is done for all concepts added to the syllabus.

At the level of the generic syllabus, the Cognitive level is considered only when following the relationships. The relationships (e.g., "ISA", "Prereq", "ComposedOf", "Follow") should be traced in OM starting at the appropriate cognition level as specified by the CLO.

The Generic Course Syllabus is adapted for each specific student to guarantee:

- No concepts/topics are not included if the student already knows them at an acceptable level.
- All concepts/topics that are pre-required for teaching the concepts of the Generic Syllabus and that are not known by the student at an acceptable level are added to adapted syllabus in order to be taught before teaching the bespoke concept. This step is recursive to capture all missing levels of the pre-required knowledge.

Therefore, the idea of the Adapted Student Course Syllabus Generator can be summarized as follows:

1) Starting at the course CLO and given the student's BDKM, the following is done to adapt the CLO into a specific student's SLO:

2) For each concept in the CLO, consult the student's BDKM, and OM:

3) If the student already knows this concept at an acceptable level of mastering, then remove it from the SLO.

4) If student's BDKM misses any of the concepts predecessors ("ISA" and "Prereq" relationships), then add this concept to the student's SLO at the same cognitive and complexity level as described in the CLO for the specified concept.

5) This last step is done recursively until is terminated by a concept that is well known to the student as per his/her BDKM.

6) Starting at the adapted SLO, do the following to generate the Adapted Student Course Syllabus:

7) For each Concept in the SLO, consult OM and his/her BDKM:

8) If the concept is known by the student, then remove it from the Adapted Syllabus.

9) Identify the concept's "ISA" and "Prereq" linked concepts, if those concepts are not known by the student's BDKM, they should be added to the Adapted Syllabus before the concept as "recall" concepts so that they are briefed to the student before start teaching the concept itself.

10) This step is done recursively to visit all pre-required concepts and their pre-requirements.

11) For each Concept, its "ComposedOf" (or sometimes called "PartsOf") relationships in OM are followed to identify the breakdown of this concept.

12) Those subordinates of the concept (its ComposedOf concepts) are ordered using the "follow" relationships among them.

13) Only subordinate concepts of complexity level less than or equal the level specified in the CLO are considered.

14) This procedure is recursively done so that the same is done for all concepts added to the syllabus.

15) The steps above are to be repeated for each time a concept is to be added to the system.

V. ADAPTIVE COURSE DELIVERY SYSTEM RECOMMENDATION

The Course Delivery System (CDS) adapts the delivery of the course to the student according to his/her student model.

The Delivery System takes it from the Adapted Student Course Syllabus, to generate the detailed Adapted Course TOC, and then to the presentation phase where the appropriate LOs are presented to the student, as shown in Fig. 4. Each student would have his/her personalized TOC. The TOC is structured into: Chapters, Sections, and Sub-Sections. Chapters and Sections come from the Adapted Student Course Syllabus. Sub-Sections are identified in this phase according to the student's LSM.

LSM Adaptation Guidelines:

The LSM adopted by the e-Learning Model is FSLSM [23] as it has applicability to e-learning and compatibility to the principles of interactive learning systems design [18]. A student's learning style will affect the adaptation process in two directions, namely, the selection and sequencing of the LOs during the course delivery.

"Selection" can be identified at large by the answers to few questions, which mainly direct the adaptation process through the selection of the appropriate LO based on the "Technical Format" attribute:

What type of information does the student preferentially perceive?

Sensory (sights, sounds, physical sensations).

Intuitive (possibilities, insights, hunches).

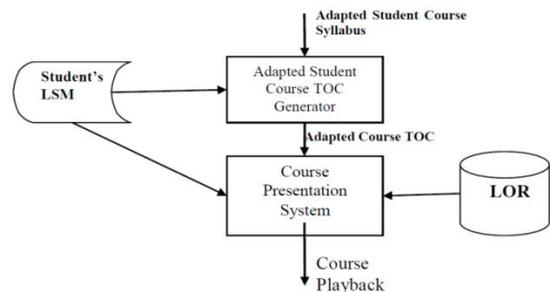


Fig. 4. The Course Delivery System Architecture

Through which sensory channel is external information most effectively perceived?

Visual (pictures, diagrams, graphs, demonstrations).

Verbal (words, sounds).

How does the student prefer to process information?

Actively (through engagement in physical activity or discussion).

Reflectively (through introspection).

How does the student progress towards understanding?

Sequentially (step by step)

Globally (in large jumps, holistically).

The guidelines governing both the Selection and Sequencing procedures are presented in [24]. Accordingly, translating these guidelines, TOC templates are designed for each of the LSM dimensions.

VI. ASSESSMENT SYSTEM RECOMMENDATION

The assessment module gathers information about the student using a test tool. LMS uses assessment tools to provide instructor with facilities to assess e-learners based on multi-type tests and exams, to track achievements in those exams and tests and to provide online grade books. The CAS displays questions from the Question Items Repository (QIR) in an adaptive way based on learner style and preferences. It is recommended to have two types of assessments: assessment after each LO presentation; and an evaluation exam for the whole presented learning section. Exam items are presented in a manner related to the learner by presenting question objects following the student's preferences. The Question Objects (QO) are parts of the question item.

The QIR is the central storage for that module, which is to be shared among instructors that maintain a collection of reusability test items to measure different levels of knowledge and skills in different difficulty levels. The CAS is adapted based on FLSM to select, present and sequence the question objects to the preferred student learning style. We are employing a simple overlay student model. It reflects the student's estimation of current knowledge levels for a student in concepts in the current domain and prerequisite concept in every level of RBT.

The instructor is responsible to identify elements of criteria for the test which are: the domain knowledge (the current course), concept to be measured and under any level of RBT (cognitive domain) wanted to measure this concept to determine the behavioral objectives, some adaptive rules related to the adaptation and evaluation process.

The engine generates the question items tailored to the student ability and based on the test objectives and instructor rules. To measure the specific level of RBT, we must measure the test objectives which are matched with this level. The grade of proficiency is set to 1 if the student has knowledge and set to 0 if the student does not have knowledge. We start to measure the objectives from a simple level to the more complex levels of RBT or vice versa depending to the concept to be measured. There are two cases, if the concept to be measured is for the current course then start from the lowest (simple) level of RBT to the required level of RBT. The other case, if the concept to be measured is for the prerequisite course, then we start from the required level of RBT (more complex) in the objectives to the lowest level of RBT.

We recommend to add assessment with many options with the following important fields:

Quiz or Exam field (Quiz or Exam) that specify if the assessment is an exam or quiz.

Question Selection field (Manual or Auto) that specify if the assessment question will be selected by the teacher (Manual) or by the system (Automatic selection) , if auto is selected then the teacher should specify number of question and their difficulty level in the fields (Number of Low Level Questions, Number of Mid-Level Questions , Number of High Level Questions)

Type field (Pre. or Post.) that specify when to view the quiz before learning object playing or after viewing it.

Concept field that shows the concept related to the assessment.

VII. DISCUSSION AND CONCLUSION

We have designed sample lectures the Web Programming course (CPIS358) at the department of Information Systems with Faculty of Computing and Information Technology at King Abdulaziz University. For web Programming course, some topics, such as JavaScript, PHP, HTML are discussed and presented based on the domain ontology prepared for the course.

The system guides the teacher throughout the course design process by helping him/her to:

- Understand the student(s) model: cognitive modes, skills, and traits;
- Determine the concepts to be covered to achieve the course learning outcomes;
- Determine the best methods and pedagogy to present those concepts to the students according to their cognitive models; and
- Search for the best available assets and learning objects that achieve such criteria.

In addition, the system helps the student during the course delivery process for the goal of making the learning process more pleasant, efficient, and effective. It will help him/her through:

- Adapt the course syllabus to match his/her background knowledge yet to meet the course objectives;
- Choose the most appropriate presentation style and pedagogy that best suits each individual student;
- Select the most appropriate course content and learning objects that suits the student the best;
- Choosing the best sequencing of the learning material;
- Identify the proper time and amount of exercises, quizzes, tests, and exams that best suites each individual student's style of learning; and
- Assessing students according to their cognitive abilities and preferences.

The following results were recognized and were conceptually proven:

- Integrating instructional design theories (e.g., RBT) and psychology and learning theories (e.g., Learning style models such as FLSM) into the adaptive learning process has been demonstrated feasible.
- Employing computer science technology to implement an intelligently adaptive authoring and delivery courses is proven feasible. Technologies such as Ontology, Learning Objects, and Expert Systems were used to achieve such goals.
- A reasonable student model was designed in such a way to achieve adaptability in delivering courses to each specific student to match his/her profile as possible for more effective and efficient self-learning process.

To the best of our knowledge, we did not find similar integrated work in our region. By carefully inspecting of some related work, we can deduce the following comparative of our system KAU-AES with other systems in the literature as shown in Table II.

For our future research directions, we may have the following points:

- Use educational data mining techniques to investigate and predict students' trends and attitude.
- Making In-Depth Analysis of the Felder-Silverman Learning Style Dimensions for our Arabic region and compare it with foreign regions.

TABLE II. COMPARISON OF ADAPTIVE SYSTEMS AND TOOLS

System	Adaptive Courses	Learning styles	Authoring Tool	SCORM Courses	LMS	Bloom Taxonomy	Ontology	Open Source
EDUCA [12]	√	Felder & Silverman	√		√			
INES [13]	√		√					
[14]			√				√	
[15]	√						√	
[17]	√		√					
[18]	√	Felder & Silverman	√					
[19]	√		√					
Moodle				√	√			√
KAU-AES	√	Felder & Silverman	√	√	√	√	√	√

ACKNOWLEDGMENT

This work was supported by King Abdulaziz City of Science and Technology (KACST) funding (Grant No. AT-204-34). We thank KACST for their financial support.

REFERENCES

[1] Chen, C. M. (2009). Personalized E-learning system with self-regulated learning assisted mechanisms for promoting learning performance. *Expert Systems with Applications*, 36(5), 8816-8829.

[2] Brusilovsky, P. (1999). Adaptive and intelligent technologies for web-based education. *KI*, 13(4), 19-25.

[3] Smith, M. K. (2003). 'Learning theory', the encyclopedia of informal education. [<http://infed.org/mobi/learning-theory-models-product-and-process/>]. Retrieved: 1-4-2018].

[4] Dunn, R., Dunn, K., & Freeley, M. E. (1984). Practical applications of the research: Responding to students' learning styles—step one. *Illinois State Research and Development Journal*, 21(1), 1–21.

[5] Dorça, F., Araújo, R., de Carvalho, V., Resende, D., & Cattelan, R. (2016). An automatic and dynamic approach for personalized recommendation of learning objects considering students learning styles: An experimental analysis. In *Informatics in education* (Vol. 15, pp. 45–62). Vilnius University.

[6] Felder, R., & Silverman, L. (1988). Learning and teaching styles in engineering education. *Engineering Education*, 78, 674–681. <http://doi.org/10.1109/FIE.2008.4720326>

[7] Kolb, D. (1984). Individuality in learning and the concept of learning styles (pp. 61–98). Englewood Cliffs, New Jersey: Prentice Hall.

[8] Mumford, A., & Honey, P. (1986). The manual of learning styles. Maidenhead, Berkshire: P. Honey, Ardingly House.

[9] Soloman, B. A., & Felder, R. M. (2005). Index of learning styles questionnaire. NC State University. Available Online at: <http://www.Engr.Ncsu.Edu/learningstyles/ilsweb.Html>. Last Visited on May 14, 2016.

[10] Sousa, D.A.: How the brain learns, 3rd edn. Corwin Press (2006)

[11] Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., Wittrock, M.: A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Complete edn. Longman (2001)

[12] Cabada, R. Z., Estrada, M. L. B., and García, C. A. R. (2011). EDUCA: A web 2.0 authoring tool for developing adaptive and intelligent tutoring systems using a Kohonen network. *Expert Systems with Applications*, 38(8), 9522-9529.

[13] Fonte, F. A. M., Burguillo, J. C., and Nistal, M. L. (2012). An intelligent tutoring module controlled by BDI agents for an e-learning platform. *Expert Systems with Applications*, 39(8), 7546-7554.

[14] F. Aparicio, M. De Buenaga, M. Rubio, and A. Hernandoet, "An intelligent information access system assisting a case based learning methodology evaluated in higher education with medical students", *Computers & Education*, Volume 58, Issue 4, May 2012, Pages 1282-1295

[15] Gladun, A., Rogushina, J., Garci, F., Martínez-Béjar, R., and Fernández-Breis, J. T. (2009). An application of intelligent techniques and semantic web technologies in e-learning environments. *Expert Systems with Applications*, 36(2), 1922-1931.

[16] Escudero, H., and Fuentes, R. (2010). Exchanging courses between different Intelligent Tutoring Systems: A generic course generation authoring tool. *Knowledge-Based Systems*, 23(8), 864-874.

[17] El-Bishouty, M. M., Chang, T. W., Lima, R., Thaha, M. B., and Graf, S. (2015). Analyzing Learner Characteristics and Courses Based on Cognitive Abilities, Learning Styles, and Context. In *Smart Learning Environments* (pp. 3-25). Springer Berlin Heidelberg.

[18] Brusilovsky, P., Sosnovsky, S., and Shcherbinina, O. (2005). User modeling in a distributed e-learning architecture. In *User Modeling 2005* (pp. 387-391). Springer Berlin Heidelberg.

[19] Gamalel-Din, S., and Al-Otaibi, R. (2008). Smart Assistant for Adaptive Course Preparation and Delivery in E-Learning Environments. In the Proceedings of the 7th European Conference on e-Learning (pp. 390-401).

[20] Al-Otaibi, R., and Gamalel-Din, S. (2010) Intelligent Querying for Adaptive Course Preparation and Delivery in E-Learning. In the Proceedings of the Ninth IASTED International Conference on Web-based Education, Sharm El-Sheikh, 15-17 March.

[21] Anderson, L. W., Krathwohl, D. R., and Bloom, B. S. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Allyn & Bacon.

[22] Felder, R. M., and Silverman, L. K. (1988). Learning and teaching styles in engineering education. *Engineering education*, 78(7), 674-681.

[23] Gamalel-Din, S., and Farida, A. S. (2008). Smart E-NoteBook: an Adaptive Hypermedia Learning Material Management Environment. In *The 7th European conference on e-learning*: [hosted by the University of Cyprus]; Grecian Bay Hotel, Agia Napia, Cyprus, 6-7 November 2008 (p. 379). Academic Conferences Limited.

[24] Farida, A. S. (2010). Student's Smart Multimedia e-Notebook. M.Sc. Thesis. Faculty of Computing and Information Technology, King Adulaziz University, Jeddah

Segmentation Method for Pathological Brain Tumor and Accurate Detection using MRI

Khurram Ejaz¹, Mohd Shafry Mohd Rahim², Amjad Rehman³, Huma Chaudhry⁴, Tanzila Saba⁵,
Anmol Ejaz⁶, Chaudhry Farhan Ej⁷

^{1,2,4} Faculty of Computing, UTM, Johor Bahru, Malaysia

³ Faculty of CCIS, Al Yamamah University, Riyadh, Saudi Arabia

⁵ Department of Info. Sys., Prince Sultan University, Riyadh, Saudi Arabia

⁶ Faculty of Allied Sci, UOL, Lahore, Pakistan

⁷ Dept. of Engineering, Nokia, R&D, Alkhobar, Saudi Arabia

Abstract—Image segmentation is challenging task in field of medical image processing. Magnetic resonance imaging is helpful to doctor for detection of human brain tumor within three sources of images (axil, corneal, sagittal). MR images are nosier and detection of brain tumor location as feature is more complicated. Level set methods have been applied but due to human interaction they are affected so appropriate contour has been generated in discontinuous regions and pathological human brain tumor portion highlighted after applying binarization, removing unessential objects; therefore contour has been generated. Then to classify tumor for segmentation hybrid Fuzzy K Mean-Self Organization Mapping (FKM-SOM) for variation of intensities is used. For improved segmented accuracy, classification has been performed, mainly features are extracted using Discrete Wavelet Transformation (DWT) then reduced using Principal Component Analysis (PCA). Thirteen features from every image of dataset have been classified for accuracy using Support Vector Machine (SVM) kernel classification (RBF, linear, polygon) so results have been achieved using evaluation parameters like Fscore, Precision, accuracy, specificity and recall.

Keywords—Brain tumor; level set; Hybrid Fuzzy K Mean (Hybrid FKM); Discrete Wavelet Transformation (DWT); Scalable Vector Machine (SVM); Magnetic Resonance Image (MRI); Principal Component Analysis (PCA)

I. INTRODUCTION

MRI gives internal visualization of soft tissues of brain and analysis if MRI is from plentiful visual information when expert when wants to examine brain for detection of brain tumor. Two kind of brain tumor have been seen in images like benign and second one is malignant. Experts check type of tumor with boundary of tissue in MRI. Three orientation of MRI are available for visualization like Sagittal (x axis), coronal (Y axis) and Axil (Z axis). In this paper Axil slice of T2 give more highlight of tumor boundary but challenge to detect due to homogenous intensity. Experts interested observing brain tumor from digital images which are noisier. To identify information from these digital images, the process of segmentation has been used. Manual segmentation is much time consuming if the volume of image becomes high. Therefore, automatic segmentation using different method becomes important and challenging for more accurate detection. Segmentation improves with combination of

thirteen texture and statistical features, reduction of features and classification to segment target labels.

In Section II detailed related work along critical table is discussed, in Section III detail analysis of two datasets is shown. Section IV is for methodology with stepwise implementation, Section V is results and discussion, and at last Section VI provides conclusion and future directions.

II. RELATED WORK

MRI imaging is compromising due to noise, complexity of detection of brain tumor disease [21]. Patient declares normal or abnormal with analysis of MR imaging. The importance of MR imaging has been seen with plentiful visualization of soft tissues [22]. MR imaging needs improvement due to signal to noise ratio [23] and they need enchantment so segmentation method diagnoses brain tumor in MR imaging and they classifies tumor as malignant or benign.

For detection of brain tumor, MR imaging incorporated by manual segmentation, semi-automatic segmentation and automatic segmentation. Manual segmentation is performed with some software's, but they have issue of variation every time. Semi-automatic segmentation has affected from personal intervention whereas automatic segmentation is incorporating with special knowledge for detection of brain tumor [24]. Special kind of filters has been used for detection of brain tumor [25]. Automatic detection is only made for local region [26]. With using combination of WFRFT+PCA+SVM achieve maximum accuracy [27].

Many studies have been found for detection of brain using classification for purpose of segmentation and results can be seen with some accurate results.

Various novel system for detection of brain tumor has used Magnetic Resonance imaging (MRI) modality [6], [15]. Using none supervised learning Fuzzy C mean, K mean for automatic segmentation are more focused for specificity and sensitivity [18], [19]. Classification techniques are more focused for accurate detection of brain tumor [1]-[6]. From methodology point of view classification is important for feature extraction [20]. Determined symmetric or none symmetric features are modelled using preprocessing and post processing techniques for pathological brain detection [6]-[10]. Feature Extraction of Brain MRI is transforming [1], In

DWT, it converts into digital values whereas in SWT can see brain features more accurately. Magnetic Resonance Brain Image has been classified based on Adaptive Chaotic PSO [2]. Combined three different techniques to find best classification for detection of normal brain or abnormal brain so applied DWT+PCA+ ADCPSO (Adaptive comparative particle swarm optimization) and after that neural network has been applied for best results. The classification results were 98.5 per cent with combination of ADCPSO+FNN over 160 images from Harvard site. The advantage of this technique was better result achievement and disadvantage of this paper is neural network working is unclear. Magnetic Resonance Brain Image Classification by AN Improved Artificial BEE Colony Algorithm for classification(ACB)[3], various classification techniques for MRI to detect brain either normal or abnormal has applied and it has been proved ACB better as compare to GA (Genetic Algorithm), Differential Evolution (DE) and PSO. For classification, it has been performed ACB (Artificial BEE colony) worse for training of FNN so it has been decided performance of optimization depend upon application, so devised Scalable Chaotic Artificial Bee Colony algorithm classifies normal and abnormal brains of T1 weight images with accuracy of 100 percent. Devised novel hybrid classifier for detection of normal and abnormal brain and has obtained 100 percent accuracy after performing experiments [4]. Before experiment has applied series of step over image like wavelet transformation at different levels then PCA for feature extraction then finally applied BNN and accurate result has found. Achieved accuracy after using combination of DWT, PCA, KSVM and kernel which has been used is GRB kernel and its accuracy is 99.38 over Axial images [5]. SVM + PSO over trained data and then optimal; optimal KSVM told about normal or abnormal brain and found 97.7 per cent accuracy which is larger amount as compare to BP-NN(86.22%), RBF-NN(91.33%) [6]. Modern classification of brain MRI images and that's that classifications are "WFRFT+PCA+GEPSVM", "WFRFT+PCA_TSVM", the proposed results are better than other eight classes[7]. Feed forward neural (FNN) had been used as classifier for optimization and in novel work has been combined with biograph based optimization and particle swarm optimization in way of training FNN, finally achieved proposed classification up to 100 percent with combination of WE(Wavelet Entropy and HBP(hybridization BO,PSO) [8]. Contribution was proved that "DWPT+TE+GEPSVM+RBF method gave better classification than other state of art methods or techniques" and accuracy was 100 per cent [9]. Successful for invention of automatic abnormal brain detection using improved classifier with combination of Q-PSO, KSVM and wavelet energy. In paper conducted experiment obtained best results comparatively. Secondly wavelet energy feature is best for abnormal brain detection. [10]. The disadvantage of this work is, this is very complex work. The good thing is that its accuracy was 97.78 percent with technique of BBO-KSVM [13]. Robustness has been defined as human visual image with abnormality and system confirms the abnormality like system did in 0.002 second to

prove robustness. Technique over 25 normal images were selected whereas rest of 25 were abnormal .90 per cent correct identification [11].

Segmentation methods are important for segmenting of MRI image and they may be fast if threshold values are assigned. Model based techniques which are based on geometric deform model they use full automatic segmentation but computationally expensive but it calculates sensitivity of tumor boundaries [12]. Performed automatic segmentation after converting knowledge into probabilistic values in image so find complex feature and Neural Network convolution is applied because it can learn more complex features in MRI therefore multi modal proposed for MRI [13]. Segmentation using morphological operation like feature of brain tumor centroid X, centroid Y, area and in comparison segmentation technique using morphological is better than other rest of methods [14]. 2D adoption noise removed, segmentation is achieved with removing strong speckle and enhance the weak boundaries of medical image, the weakness is accomplished with range filter to segment various anatomical structure [15]. In first generation segmentation methods are threshold methods, region growing methods and edge-based detection and in second generation they are cluster, classifier, deformation model and graph cluster. In third generation, graph guided approach, shape model, appearance model, medical image segmentation methods, algorithm and application. [16]. Segmentation methods are important for detection of brain tumor when MRI is compromising different issues during their process. Combined K mean cluster algorithm with fuzzy C scan in minimal execution time in four stages like preprocessing, clustering, feature extraction and validation from noisy portion of MRI [17]. An unsupervised method with a clustering approach for tumor identification and tissues segmentation in magnetic resonance is important from an unsupervised perspective [18]. Two techniques hybrid Fuzzy c mean, k mean using SOM achieved best sensitivity, mean square root, specificity over Harvard repository images and has achieved best segmentation over three sources of image like axial images, coronal and sagittal [28]. Vector classification and Z indexing [19] is good paper for classification point of view of classification and feature can be Efficient Feature detection of image using Multimedia database using Query [20]. Contextual, fuzzy classification methods, robust features and extraction methods achieve good results during segmentation [32]-[36].

The segmentation of brain tumor has been performed in continuous regions as well as in discontinuous regions. Brain tumor is as a feature and has been detected in MRI when applies quad tree for detection of region of interest (ROI) in continuous region [21]. Quad tree active contour level set are compromising in discontinuous regions. Hybrid FKM-SOM technique has filled gap of partitioning of tumor portion and edema region, but limitation occurs due to intensities mixing, analyzed three sources of images like Axial, Sagittal and Coronal but segmentation results using comparison parameter are not sufficiently proving for segmentation and evolution of brain tumor.

TABLE I. ACCURACIES OF SEGMENTATION

Paper title	Techniques	Segmentation Accuracy
(Sneha Dhurkunde 2016)	Histogram, Fuzzy c mean, K mean	79.5
(Saleha Masood*, 2015)	Thresholding, Region growing, Clustering, Classifiers, Bayesian approach, Deformable methods, atlas guided approach, edge based methods, compression based method	Missing
(Norouzi <i>et al.</i> , 2014)	Thresholding, histogram, Region of interest, Clustering techniques, Classification techniques, Expectation maximization, Graph Cut	Missing
(Rajaei <i>et al.</i> , 2012)	Image texturing, Range filters	Missing
(Padma and Sukanesh, 2011)	Dominant grey level run length	85
Vishnuvarthanan <i>et al.</i> , 2016).	Hybrid Fuzzy k mean-SOM, Fuzzy c mean, neural network, evaluation matrix	91
(Abdel-Maksoud <i>et al.</i> , 2015)	Median filter, K mean cluster, Fuzzy c mean	87.5
Pei, L., Reza, S. M., and Iftekharuddin, K. M. (2015)	K cluster, histogram, joint label fusion	71
Yang, G., Zhang, Y., Yang, J., Ji, G., Dong, Z., Wang, S., et al. (2016)	Classification, Pattern Recognition, Support Vector Machine, Biogeograph based optimisation	Missing
Zhang, Y., Dong, Z., Wang, S., Ji, G., and Yang, J. (2015)	Shannon entropy; Tsallis entropy; discrete wavelet packet transform, support vector machine kernel technique, pattern recognition; classification	Missing
(Sudharani <i>et al.</i> , 2016)	Sampling, histogram, morphological operation	89

Issues are arisen in brain tumor using MRI using segmentation when brain spread in discontinuous regions, due human less interactivity especially when MRI is nosier so to separate tumor tissues from normal tissues with accuracy become challenge. Accuracy of with combination of different segmentation method can be seen in Table I.

III. MATERIAL

A. Data Set

Fig. 1 gives view of Scope of work from dataset no. 1, Two datasets have been based analyzed like Harvard data set and one local hospital MRI data set. Data set 1 is based on

three orientation for single patient of MRI. Dataset number 1 is consisting clinical used MR images dataset and it is consisting of eleven total patients; patient class has been divided into two main classes and their names are malignant or Benign. According to history of patient we have seven malignant patients and four Benign patients. Both classes are for of abnormal patients. One patient is further divided into three kind of images plane like corneal (y axis plan), sagittal (x axis plane) and z axis plane which is known as axil plane. Dataset is consisting of T2 sequences. T2 sequence images are more enhance so that radiologist can easily observe before operation and treatment

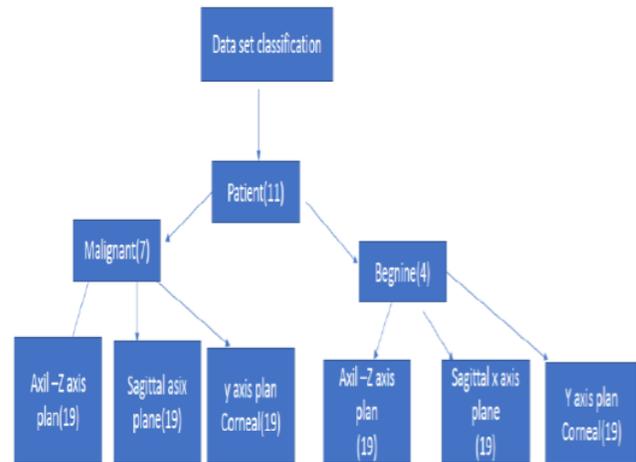


Fig. 1. Proposed feature diagram for data set number

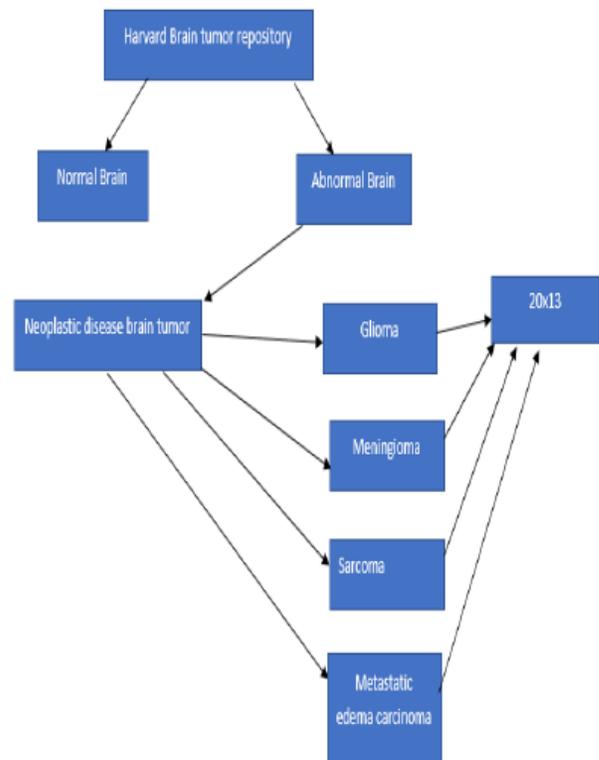


Fig. 2. Data set number 2 diagram

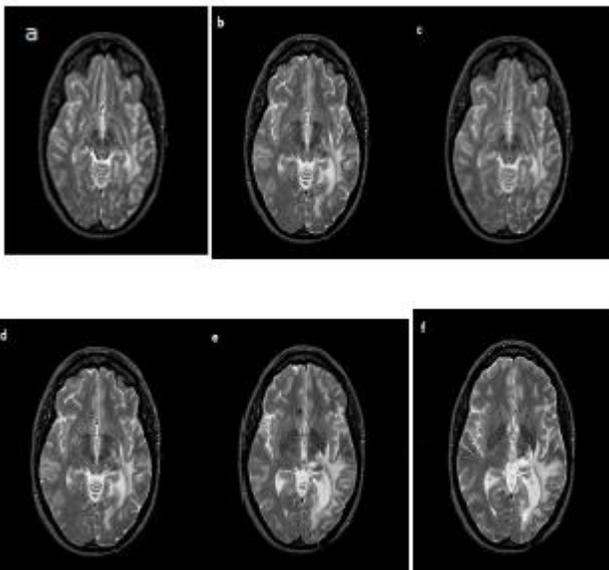
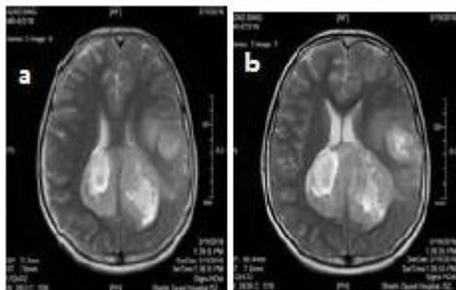


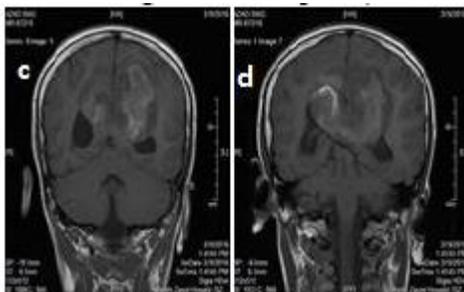
Fig. 3. Harvard dataset used by Y Zhang dataset (2).

Fig. 2 is from benchmark dataset from Harvard brain published repository. The bench mark paper used Harvard repository of axial images [5]. Same wise [5], same orientation of dataset used same by [7] and are performing experiment over axil Z-axis images for accurate identification of brain tumor. These images are MR images and novel accurate has been evaluated for identification of brain tumor so it's one view of images and for these images output accuracy of classification is more than 90 per cent. Other planes are compromising. In data set result of classification is not mentioned.

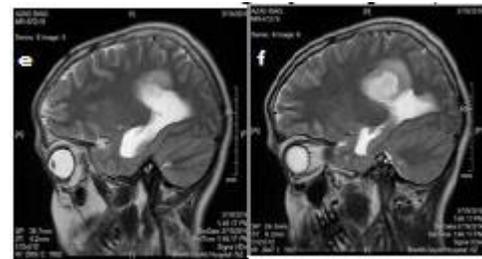
Datasets images has been taken for same patient, all of axis plane data set like in Fig (a), (b), (c), (d), (e), and (f) for same patient.



(Dataset 1 Axil image for x axis plane)



(Dataset 1 Corneal images y axis plane)



(Dataset 1 sagittal x axis plane)

Fig. 4. Fig. a, Fig. b, Fig. c, Fig. d, Fig. e, Fig. f for x, y, z-axis plane.

Above Fig. 3 and 4 are from two datasets like clinical dataset and another Harvard published dataset.

IV. METHODOLOGY

The methodology in Fig. 5 shows ten steps, and Fig. 6, 7, 8 for contour and Fig. 9 and 10 for DWT and Fig. 11 for SVM classification. In first step dataset has been input, in step two dataset has been preprocessed for noised reduction, contour has been initiated in step three for detection of tumor location, in step number 4 segmented portion achieved, in step number 5 features of image using DWT have been extracted for accurate detection, in step number sixth achieved features have been reduced using PCA, in step number seventh 13 features have been achieved, in step number eight flavor of KSVM(Linear, polygon) achieved accuracy of classification and at second last step evaluation of accuracy achieved through specificity, accuracy, Fscore, precision and Recall have been derived and last one are results so comparison of dataset can be achieved.

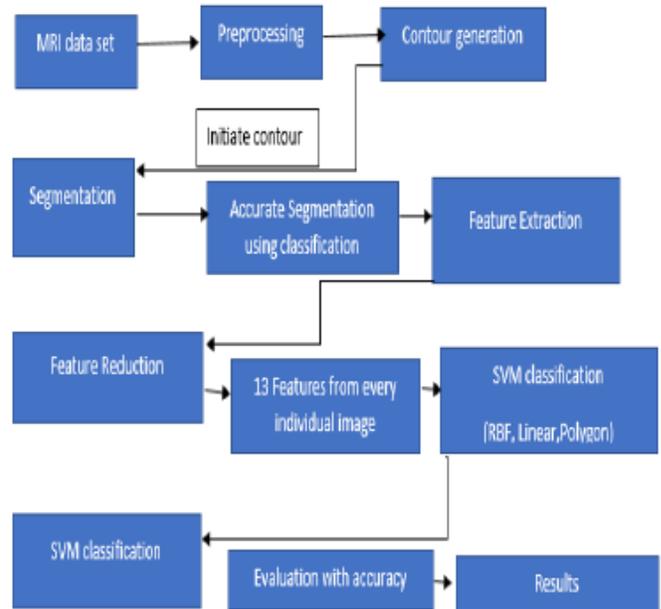


Fig. 5. Frame work

Image segmentation performed with level set function the contour level zero so has been chosen manually [32]. Segmentation done without selection of proper initiate contour (IC) effects the segmentation results for discontinuous tumor within MRI. Therefore, in proposed strategy proper contour

has been generated, firstly binarization [31] has been performed seconding remove an essential then contour map have been generated for detection of discontinuous tumor portion as feature discontinuously region of interest. From Fig. 6 and 7 contour map has been generated.



Fig. 6. Original image DS1.



Fig. 7. Binarization Fig. (6).



Fig. 8. Initiated Contour map generated

DWT (Discrete Wavelet transformation) is applying, image is considering as a signal and presents in time domain. [29] In time domain image has been divided locally into bands so energy level of image features has been highlighted.

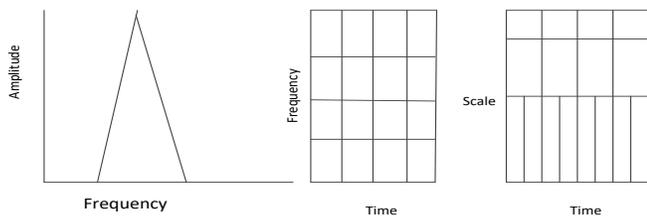


Fig. 9. DWT signal representation

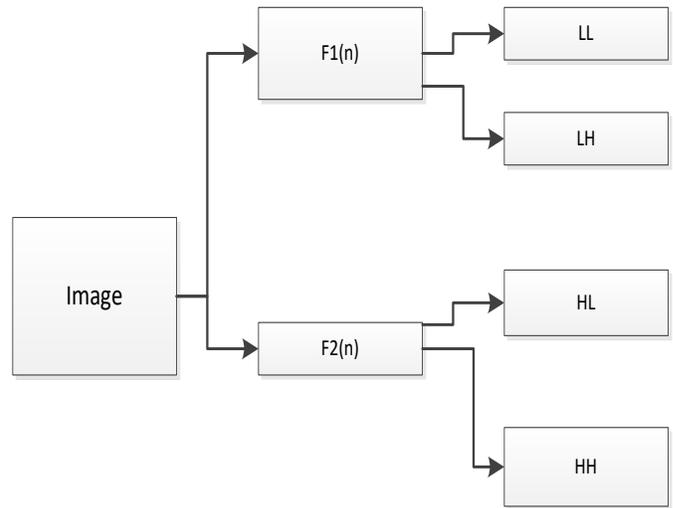


Fig. 10. DWT signal representation with high pass and low pass variations.

In step number 6, PCA (Principal Component Analysis) performs for dimension reduction of image features. Firstly, identifies uncorrelated elements from image then least uncorrelated element remove from data set images so the group of most correlated elements have been obtained. Fig. 9 to 10 shows DWT representation of image.

Step number 8 is accurate feature classification, obtained elements are classified through Support Vector Machine(SVM) like two classes are [-1,1]. SVM gives hyper plane and in such plan classified output(relevant) classes of images in query [30].

$$F(x, y) = \{(x_n, y_n) / x_n \in R^p, y_n \in (1, -1)\}$$

To address above two issues firstly I have tested my dataset which is only Axil (z axis images) image so I could see one picture of dataset then I have trained my dataset using DWT, PCA, SVM and Kernel (Fig. 11) based SVM and next I have proposed dataset with aim of accurate classification of all of three kind images.

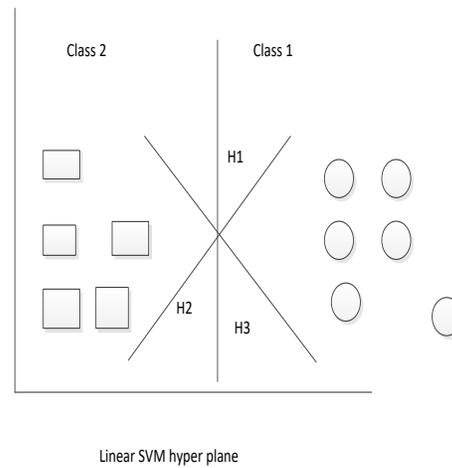


Fig. 11. KSVM has variation of SVM like RBF kernel, quadratic kernel and linear kernel for accurate detection of features in queried image.

TABLE II. REFERENCED BASED 13 FEATURES TABLE

Feature	Reference
contrast	Zhang, Y., and Wu, L. (2012)
corelation	Zhang, Y., Ji, G., Yang, J., Wang, S., Dong, Z., Phillips, P., et al. (2016)
energy	Zhang, Y., and Wu, L. (2012)
homogeneity	Zhang, Y., Wang, S., Dong, Z., Phillip, P., Ji, G., and Yang, J. (2015b)
mean	(Abdel-Maksoud <i>et al.</i> , 2015), (Vishnuvarthanan <i>et al.</i> , 2016)
standard deviation	S., Ji, G., and Yang, J. (2015)
root mean square	(Abdel-Maksoud <i>et al.</i> , 2015), (Vishnuvarthanan <i>et al.</i> , 2016)
variance	S., Ji, G., and Yang, J. (2015)
kurtosis	S., Ji, G., and Yang, J. (2015)
skewness	Zhang, Y., and Wu, L. (2012)
inverse difference movement	Zhang, Y., Ji, G., Yang, J., Wang, S., Dong, Z., Phillips, P., et al. (2016)

Above Table II is list of features which we have used for accurate segmentation.

The pseudo code of our work is in nine steps and state flow In Fig. 12, Step wise construction can be seen for segmentation of tumor with methods and accuracy has been checked with flavor of kernels.

1. Input n MRI brain tumor dataset
2. Preprocess data using (Axil, Cornea, Sagittal)
3. Contour has been generated for find segmentation
4. Segmented portion achieved
 - 4.1 classification improvement using piece wise function for measuring none uniform intensities and HOG(histogram orientation gradient) for feature of image.
5. Feature extraction through DWT
6. Features reduction through PCA
7. 13 Features achieved across n datasets
8. Segmentation classification achieve through KSVM (RBF, Linear, Polygon)
9. Improved comparative parameters values achieved
 - 9.1 Segmented kernel-based accuracy value
 - 9.2 Segmented kernel-based Fscore value
 - 9.3 Segmented kernel-based Specificity value
 - 9.4 Segmented kernel-based Precision value
 - 9.5 Segmented kernel-based Recall value
10. Results

Fig. 12. Segmentation with step wise construction.

V. RESULTS AND DISCUSSION

In this paper, two data sets have been checked. Dataset number 1 is taken from sheikhzaidh hospital and second data set MR images have been taken from Harvard medical school repository.

Data set number 1, which is consisting of MR images, MR images are three sources like Axil, Corneal and Sagittal. For axil in Fig. 13 and Table III, corneal in Fig. 14 and Table IV and sagittal with comparison to bench mark in Table V and Fig. 15. In dataset number 1, eleven patients have been checked and for each patient 99x13(axil), 99x13(sagittal) and 99x13(corneal) has been overlooked and detected segmented tumor with accuracy and RBF, linear and polygon kernels whereas dataset number 2 which is consisting of only axil images for 42 years old patient. Same 13 features have been selected so total axil images are 20x13. The target labels are 0 or 1. Zero mean malignant however 1 means benign. Same linear, polygon kernel, RBF and linear have been used for accurate detection. Fig. 12 and Table II are from bench mark dataset.

In this section using SVM three kernels, their names are polygon kernel, RBF and linear and their results can be seen kernel versus comparison evaluation parameters like accuracy, specificity, precision, recall and FScore. Experiment is performed over clinical dataset and over public data. In Fig. 13 a graph figure, benchmark dataset accuracy of segmentation can be seen across evaluation metrics. Fig. 14 is giving evaluation of Axil orientation segmentation accuracy whereas Fig. 15 gives picture of accuracy of coronal images. Fig. 16 the last one is showing picture of accuracy of sagittal along comparative analysis among these three orientations.

TABLE III. HARVARD BENCH MARK DATASET DATA SET NUMBER2 USING RBF, LINEAR AND LINER KERNEL

	Zhong et al Accuracy	Specificity	Precision	Recall	Fscore
Linear kernel	66.6666667	66.66666667	66.66667	66.66667	66.66667
RBF kernel	50	33.33333333	66.66667	50	57.14286
Polynomial Kernel	33.3333333	66.66666667	0	0	0

Table III is targeting Harvard data set and three kernels (Linear, RBF, Polynomial) have been checked across for evaluation.

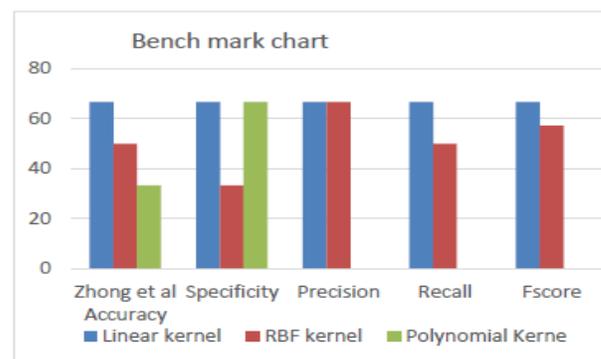


Fig. 13. Comparative chart

TABLE IV. DATASET NUMBER 1 AXIL IMAGES ACCURACY

Accuracy of Axil images with Dataset1					
	Accuracy SZ	Specificity	Precision	Recall	Fscore
Linear	66.666	95	10	50	16.667
RBF	73.33	100	20	100	33.33
Polynomial	73.33	80	60	60	60

Table IV is depict accuracy of clinical dataset number I across evaluation parameters.

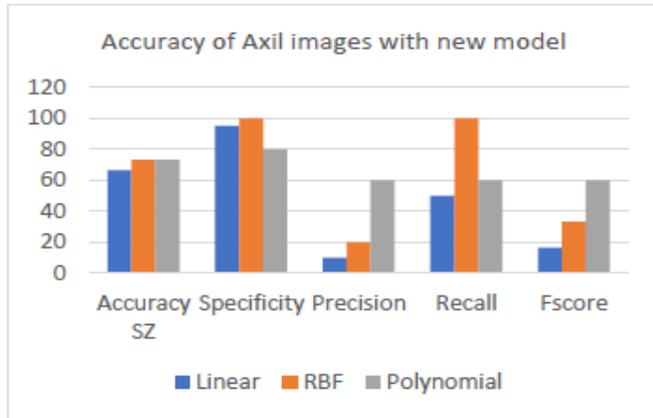


Fig. 14. Graph for axil images from dataset number 1

TABLE V. DATASET NUMBER 1 ACCURACY RESULTS WITH RBF, POLYGON AND LINEAR KERNEL

Accuracy of Corneal images with Dataset no. 1					
	Accuracy SD corneal	Specificity	Precision	Recall	Fscore
Linear	80	100	40	100	57.14286
RBF	96.66667	95	100	90.90909	95.2381
Polynomial	90	85	100	76.92308	86.95652

Table V is giving accuracy of orientation of coronal using three kernels for accurate segmentation of tumor.

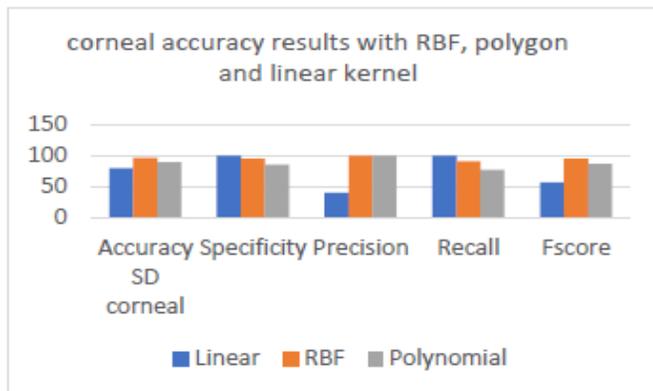


Fig. 15. Graph for Coronal images from dataset number 1

TABLE VI. TABLE FOR SAGITTAL IMAGES FROM DATASET NUMBER NO. 1 AND COMPARISON ACCURACY OF BOTH DATASETS

Accuracy for comparative analysis of Dataset1 with bench mark dataset					
	Accuracy SZ Sagittal	Specificity	Precision	Recall	Fscore
Linear	40	68.75	7.142857143	16.66666667	10
RBF	53.33333333	100	0	NaN	0
Polynomial	46.66666667	43.75	50	43.75	46.66667

Table VI is for accuracy of kernels over dataset number 1 of sagittal (X axis).

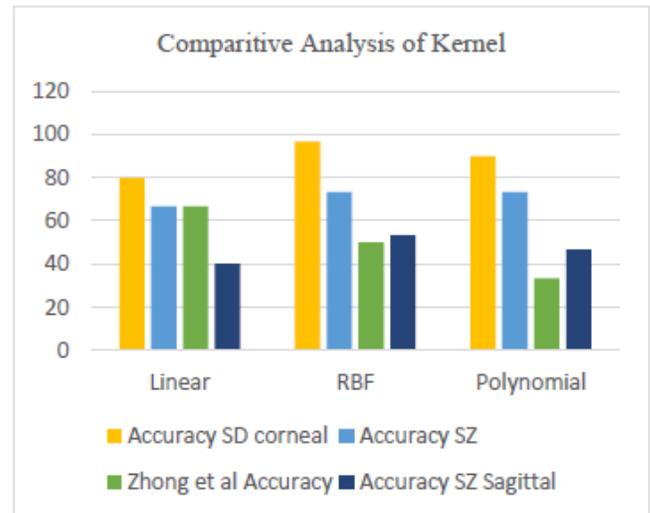


Fig. 16. For comparison of accuracy of both data set number 1 and dataset number 2

From above graph we can say clear cut difference of two dataset value for accurate segment detection of brain tumor using RBF, Linear and polynomial kernel. More transparent comparison has been drawn among datasets in above graph. Using linear kernel both datasets are equal whereas using RBF dataset 1 values need improvement and if using polynomial kernel then dataset 2 need more improvement as compare to dataset number 1.

VI. CONCLUSION AND FUTURE WORK

We have identified new improved scheme of steps of segmentation like appropriate contour generation, hybrid SOM-FKM for identification of tumor in discontinuous region and segmentation accurate results have been highlighted with evaluation parameters like accuracy, Fscore, precision and recall using RBF and polygon kernel where linear accuracy of dataset are almost 66.6 per cent whereas using RBF Harvard dataset used by Y Zhang accuracy is less than using dataset 1 so need to improve accuracy. In this paper, analyzed three sources of images (axil, corneal, sagittal) in first dataset whereas in second dataset (Harvard) has been only using axil images therefore future work will also to analyzed rest of source of images in dataset 2.

REFERENCES

- [1] Zhang, Y., et al. Feature extraction of brain MRI by stationary wavelet transform. in Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on. 2010. IEEE.
- [2] Zhang, Y., S. Wang, and L. Wu, A novel method for magnetic resonance brain image classification based on adaptive chaotic PSO. Progress In Electromagnetics Research, 2010. **109**: p. 325-343.
- [3] Zhang, Y., L. Wu, and S. Wang, Magnetic resonance brain image classification by an improved artificial bee colony algorithm. Progress In Electromagnetics Research, 2011. **116**: p. 65-79.
- [4] Zhang, Y., A hybrid method for MRI brain image classification. Expert Systems with Applications, 2011. **38**(8): p. 10049-10053.
- [5] Zhang, Y. and L. Wu, An MR brain images classifier via principal component analysis and kernel support vector machine. Progress In Electromagnetics Research, 2012. **130**: p. 369-388Y.
- [6] Zhang, Y., et al., An MR brain images classifier system via particle swarm optimization and kernel support vector machine. The Scientific World Journal, 2013. 2013.
- [7] Zhang, Y.D., et al., *Magnetic resonance brain image classification based on weighted-type fractional Fourier transform and nonparallel support vector machine*. International Journal of Imaging Systems and Technology, 2015. **25**(4): p. 317-327.
- [8] Zhang, Y., et al., Pathological brain detection in magnetic resonance imaging scanning by wavelet entropy and hybridization of biogeography-based optimization and particle swarm optimization. Progress In Electromagnetics Research, 2015. **152**: p. 41-58.
- [9] Zhang, Y., et al., *Preclinical diagnosis of magnetic resonance (MR) brain images via discrete wavelet packet transform with Tsallis entropy and generalized eigenvalue proximal support vector machine (GEPSSVM)*. Entropy, 2015. **17**(4): p. 1795-1813.
- [10] Zhang, Y., et al., *Preliminary research on abnormal brain detection by wavelet-energy and quantum-behaved PSO*. Technology and Health Care, 2016. **24**(s2): p. S641-S649.
- [11] Gordillo, N., E. Montseny, and P. Sobrevilla, State of the art survey on MRI brain tumor segmentation. Magnetic resonance imaging, 2013. **31**(8): 1426-1438.
- [12] Yang, G., et al., Automated classification of brain images using wavelet-energy and biogeography-based optimization. Multimedia Tools and Applications, 2016. **75**(23): p. 15601-15617
- [13] Dimililer, K. and A. İlhan, Effect of Image Enhancement on MRI Brain Images with Neural Networks. Procedia Computer Science, 2016. **102**: p. 39-44.
- [14] Işın, A., C. Direkoğlu, and M. Şah, Review of mri-based brain tumor image segmentation using deep learning methods. Procedia Computer Science, 2016. **102**: p. 317-324.
- [15] 1, S.D., Segmentation of Brain Tumor in Magnetic Resonance Images using Various Techniques. International Journal of Innovative Research in Science, Engineering and Technology, 2016.
- [16] Rajaei, A., L. Rangarajan, and E. Dallalzadeh, Medical Image Texture Segmentation Usingrange Filter. computer science and information technology, 2012. **2**(1).
- [17] Masood, S., et al., A survey on medical image segmentation. Current Medical Imaging Reviews, 2015. **11**(1): p. 3-14.
- [18] Abdel-Maksoud, E., M. Elmogy, and R. Al-Awadi, Brain tumor segmentation based on a hybrid clustering technique. Egyptian Informatics Journal, 2015. **16**(1): p. 71-81.
- [19] Vishnuvarthanan, G., et al., An unsupervised learning method with a clustering approach for tumor identification and tissue segmentation in magnetic resonance brain images. Applied Soft Computing, 2016. **38**: p. 190-212.
- [20] Ejaz, K., A. Mateen, and I. Ehsan. Vector Shape Classification and Z indexing. in IEEE Intl. Conference on Intelligent Network and Computing ICINC. 2010.
- [21] Kumar, E.P., Kumar, V.M. and Sumithra, M.G., 2013, July. Tumour detection in brain MRI using improved segmentation algorithm. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-7). IEEE.
- [22] S. K. Nayak, Y. K., Dr. C. S. Panda. (2015). A Study on Brain Mri Image Segmentation Techniques International Journal of Research Studies in Computer Science and Engineering Volume 2, , PP 4-13
- [23] Saleha Masood*, M. S., Afifa Masood, Mussarat Yasmin and Mudassar Raza (2015). A Survey on Medical Image Segmentation. Current Medical Imaging Reviews(Segmentation in medical imaging), 3-14
- [24] Gordillo, N., Montseny, E., and Sobrevilla, P. (2013). State of the art survey on MRI brain tumor segmentation. Magnetic resonance imaging, **31**(8), 1426-1438.
- [25] Rajaei, A., Rangarajan, L., and Dallalzadeh, E. (2012). Medical Image Texture Segmentation Usingrange Filter. computer science and information technology, **2**(1).
- [26] Sudharani, K., Sarma, T., and Prasad, K. S. (2016). Advanced morphological technique for automatic brain tumor detection and evaluation of statistical parameters. Procedia Technology, **24**, 1374-1387.
- [27] Gonzalez, R. C., and Woods, R. E. (2005). Book on "Digital image processing": Prentice-Hall of India Pvt. Ltd.
- [28] Vishnuvarthanan, G., Rajasekaran, M. P., Subbaraj, P., and Vishnuvarthanan, A. (2016). An unsupervised learning method with a clustering approach for tumor identification and tissue segmentation in magnetic resonance brain images. Applied Soft Computing, **38**, 190-212.
- [29] Zhang, Y.D., et al., Magnetic resonance brain image classification based on weighted-type fractional Fourier transform and nonparallel support vector machine. International Journal of Imaging Systems and Technology, 2015. **25**(4): p. 317-327.
- [30] Yang, G., et al., Automated classification of brain images using wavelet-energy and biogeography-based optimization. Multimedia Tools and Applications, 2016. **75**(23): p. 15601-15617.
- [31] Lopes, N.V., do Couto, P.A.M., Bustince, H. and Melo-Pinto, P., 2010. Automatic histogram threshold using fuzzy measures. IEEE Transactions on Image Processing, **19**(1), pp.199-204.
- [32] Gao, H. and Chae, O., 2010. Individual tooth segmentation from CT images using level set method with shape and intensity prior. Pattern Recognition, **43**(7), pp.2406-2417.
- [33] Lung, J.W.J., et al., Fuzzy phoneme classification using multi-speaker vocal tract length normalization. IETE Technical Review, 2014. **31**(2): p. 128-136.
- [34] Sharifara, A., M.S.M. Rahim, and M. Bashardoost. A novel approach to enhance robustness in digital image watermarking using multiple bit-planes of intermediate significant bits. in Informatics and Creative Multimedia (ICICM), 2013 International Conference on. 2013. IEEE.
- [35] Jabal, M.F.A., et al. A comparative study on extraction and recognition method of CAD data from CAD drawings. in Information Management and Engineering, 2009. ICIME'09. International Conference on. 2009. IEEE.
- [36] Harouni, M., et al., Online Persian/Arabic script classification without contextual information. The Imaging Science Journal, 2014. **62**(8): p. 437-448.

Skew Detection and Correction of Mushaf Al-Quran Script using Hough Transform

Salem Saleh Bafjaish¹, Mohd Sanusi Azmi²,
Mohammed Nasser Al-Mhiqani³,
Amirul Ramzani Radzid⁴

Faculty of Information and Communications Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Hairulnizam Mahdin⁵

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Johor, Malaysia

Abstract—Document skew detection and correction is mainly one of base preprocessing steps in the document analysis. Correction of the skewed scanned images is critical because it has a direct impact on image quality. In this paper, the authors proposed a method for skew detection and correction for Mushaf Al-Quran image pages based on Hough transform method. The technique uses Hough transform lines detection for calculating the skew angulation. It works for different version of Mushaf Al-Quran image pages which has skewed text zones. Moreover, it can detect and correct the skew angle in the range between 20 degrees. Experiment conducted on different Mushaf Al-Quran image pages shows the accuracy of the method.

Keywords—Skew detection; skew correction; Hough transform; preprocessing; binarization; image analysis

I. INTRODUCTION

Document Image processing is one of the fields that are rapidly growing faster in nowadays. It aims to convert paper-based documents to forms that are proper for storage. It can be defined as the method that is used to perform some operation on specified image such as (Digitization, Storage, compression, Re-printing) [1]. Besides that, there are different aspects that image processing could be the base such as, electronic engineering and computer science too. One of the problems in this field is that, the text in a document may be rotated when scanning which leads to produce a skewed text in the printed document as in Fig. 1.



Fig. 1. Al-Quran Surah Al-Fatiha with Skew Angle -8°.

As a result of that, the quality of the document is decreased and that will lead to multiple problems in analysis the image as well as reduce performance of optical character recognition (OCR) [18]. This paper focuses on skew detection and correction for Mushaf Al-Quran image pages. By comparison to other language scripts, skew detection and correction for Mushaf Al-Quran script is quite different as it has diacritical marks as well as the handwritten style is different too as compared to normal Arabic scrips. Hough transform is a simple feature extraction technique that is widely used in computer vision, image analysis and image processing as well. It can simply use to find lines in image by linear transform to detect straight lines [9].

II. RELATED WORK

Many studies of skew correction are published but for different languages such as English, Urdu, Chinese. However, in the document, text can be written on several text lines. A various methods are used for skew detection and correction based on different algorithm like, Projection profile, nearest neighbor clustering, Fourier transform, cross correlation and others. Skew can be defined as the angle that deviates from x-axis. Furthermore, accurate skew detection and correction helps other processes of OCR to be more successful. In [2] a novel method was proposed to recognize Arab / Jawi and roman digit by OCR. In [3] skew in documents can be classified into three class namely global skew, multiple skew and no-uniform text line skew. In [4] document analysis depends on preprocessing stage, the much better the image is preprocessed, a much better result of analysis the image is. Furthermore, it increases the quality and the accuracies in the OCR systems. In [5] skew detection and correction can be the first step in the process of the document analysis as well as understanding processing steps as it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages. Currently, a lot research in Arabic documents but less work is intensively been explored for Mushaf Al-Quran. Initially method to estimate the skew angle in a paper as in Fig. 2 is to draw a line through the text characters, and then the angle of the drawn line with the horizontal edges of the original paper is the skew angle.

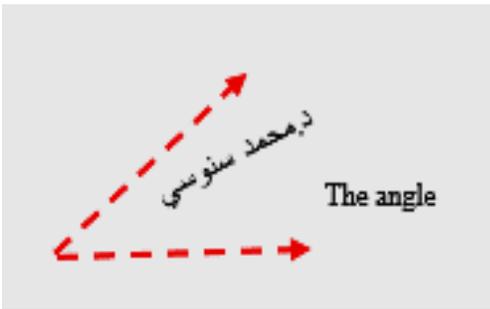


Fig. 2. Basic Skew Angle.

Generally, all ordinary pages have the skew angle of zero. However, the skew angle occurs due to different reasons. The main purpose of skew detection and correction is required to improve the quality of the scanned documents. In [6] O’Gorman paper, all these techniques can be categories into three groups as projection profile, Hough transform and nearest neighbor clustering. In [3] an evaluation for the most frequently skew detection techniques cited in their paper as (i) Projection Profile Analysis (PP), (ii) Hough Transform (HT) and (iii) Nearest Neighbor(NN). A comparison between the three techniques, the comparison started the weakness and strengths of each method as well as to compare the performance for both of them in term of the speed and the accuracy. Their evaluation showed that nearest neighbor techniques is the fastest one among them according to the speed but in other hand, its accuracy estimation evaluation is poor comparing to the other techniques. Furthermore, project profile technique gave the best estimation for the angle when it comes to the accuracy, in opposite its time is the longest to be executed. In [7] an efficiency discussion of two techniques Principal Component analysis (PCA) and Hough transform is presented to overcome problems that spoils the scanned documents. In [8] projection profile method is proposed for skew detection for handwritten signature, they used horizontal projection for detecting the skew angle and correct it using rotation transformation. In [9] a method proposed for detecting the skew and correct it for the handwritten Devanagari script using the technique Hough transform. The proposed method is to detect the skew and correct it at the word level as Devanagari script as in Fig. 3 is a little difficult comparing to other scripts because of the style of the writing as well as the writing style differs from one person to the other one [9].

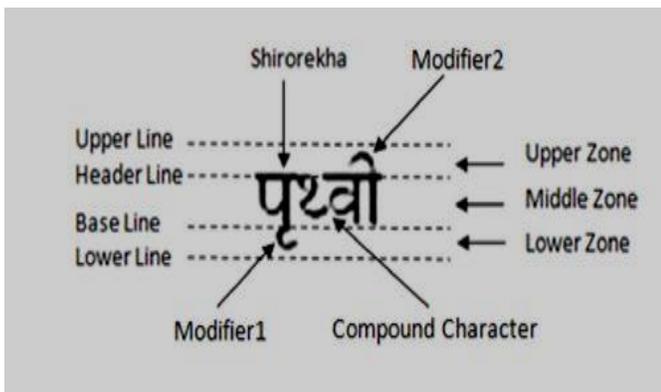


Fig. 3. Devanagari Simple Character [9].

The proposed method consists of preprocessing stage followed by word extraction stage is made in the image in order to extract the words, lastly Hough transform algorithm is applied in order to detect the skew of the word. In [4] proposed a novel skew detection and correction approach for scanned documents contains of two stages, first find the angles of the lines in the image with the respect of x-axis and second find the exact skew angle from the angles that are extracted from lines in the first stage. In [11] a proposed a simple and fast algorithm that determine the skew angle of the image as well as the slant angle of the text characters using the gradient orientation histogram. Additionally, the angle can be obtained using searching for a peak in the image histogram, the image can be corrected by a rotation at such an angle. In [1] proposed a new technique that detect the skew and correct it for the Arabic printed scripts based on connected component analysis and pixel projection. Moreover, the proposed technique take the advantage of the sharp writing line property for Arabic language that is obtained from histogram projection of the image for skew detection. In [12] an image moments are used for skew detection and correction. An image moment is the calculation of the weighted average (Moment) of the pixels 'intensities of the image. So, moments are employed to find the primary axis of every object in the document instead of applying the Hough transform. Finally, by using a feature that depends on the size of the object, the weighted average angle is estimated. In [13] skew detection method that uses run-length and Hough transform algorithm is presented. The proposed method reduce the amount of data in the image through using black horizontal and vertical run-lengths histograms which also reduces computational calculation of Hough transform and increase the speed of skew detection.

III. MUSHAF AL-QURAN SCRIPTS CHALLENGES

Mushaf Al-Quran is the holy book millions of Muslims around the world. It can be in two versions digital or printed form, although is in Arabic, but the way it is written is different from any Arabic/Jawi based document as it has “diacritics”. In [14] a proposed method for identifying types of Arabic calligraphy in Malay accent script that is written in Jawi. Fig. 4 illustrates most of Arabic script challenges as in [15], [16] have presented Arabic scripts challenges as the following:

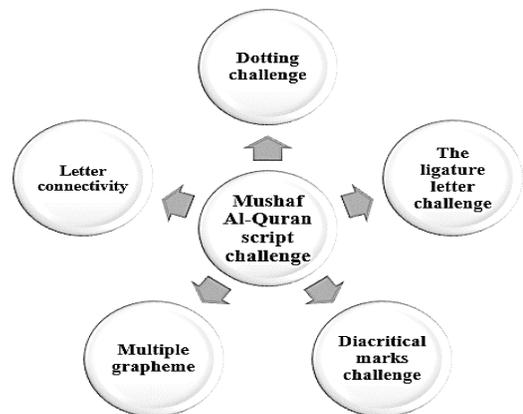


Fig. 4. Arabic Script Challenges.

1) *The connectivity challenge*: Arabic text can be only scripted cursively, that means all graphemes are connected together, this happens whether the text was handwritten or font written as in Fig. 5.

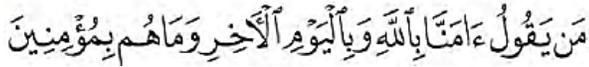


Fig. 5. Al-Quran Surah Al-Baqarah 8.

2) *The dotting challenge*: Dots in Arabic scripts are used to differ between the characters sharing similar graphemes. Accordingly, if a dot is missed with the process of skew detection, then that will affect the meaning of the text. Fig. 6 illustrates the dotting challenge.



Fig. 6. Dotting Letters.

3) *The multiple grapheme cases challenge*: In Arabic orthography it's very due to have the connectivity in text which means that same letter can be different in the way how it's written based on the position of it in the Arabic word. Fig. 7 illustrates the letter ع with different writing styles.

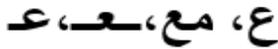


Fig. 7. Multiple Grapheme.

4) *The ligatures challenge*: Character in Arabic script can be compounded together at certain positions of the Arabic word. Ligatures can be found at almost all the Arabic fonts. Fig. 8 illustrates ligatures challenge.

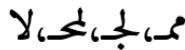


Fig. 8. Ligature Letters.

5) *The diacritics challenge*: The usage of diacritical marks helps to resolve linguistic ambiguity of the text. [14] However, in some case they goes vertical while the main text is going straight on line (horizontal) from right to left. Therefore, that makes some confusion for skew detection step in OCR. Fig. 9 illustrates a segment of Mushaf Al-Quran text with diacritics marks.



Fig. 9. Al-Quran Surah Al-Hujurat 29.

IV. PROPOSED METHOD

The proposed methodology for Mushaf Al-Quran skew detection and correction is described here. The proposed method consists of six stages namely as convert to grayscale image, binary image, foreground image, Hough transform method to detect lines, calculate skew angle and finally rotate image as in Fig. 10.

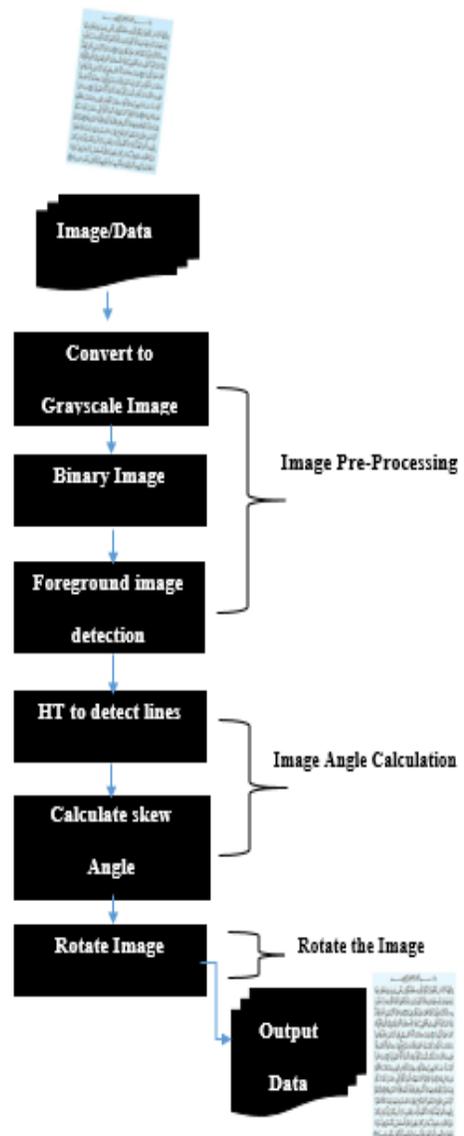


Fig. 10. Proposed Method.

A. Grayscale Image

As in Fig. 11 some of Mushaf Al-Quran pages comes with different colors, so there is a need to re therefore there is a need for the conversion to grayscale image to get high performance of skew detection. There are some reasons for converting color Mushaf Al-Quran images to grayscale images as the following:

- 1) Reduce color: in color images, sometimes information of the images doesn't help to identify the important areas on the images and other features such as lines on the images.
- 2) Grayscale (8bits) images makes it easy for implementing binary algorithms because there are only two shades of colors in grayscale images which are white and black whereas color images has blue- green-red.
- 3) Algorithms applied to grayscale images are much faster than the once applied to RGB images.

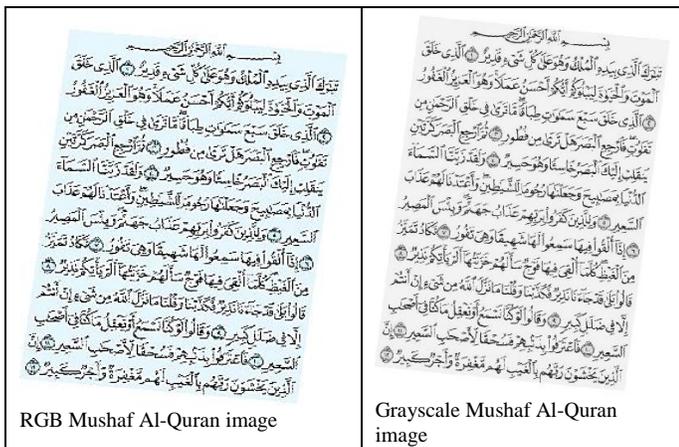


Fig. 11. Al-Quran Surah AlFajr (Different between color and grayscale image of Mushaf Al-Quran).

B. Binarization

In [17] an amendment has been made by applying Otsu's method for to improve noise and prepare images for the new proposed extraction feature method. In [10] also Otsu's method is applied for Arabic characters dynamically in order to choose the discriminant threshold on the image. Therefore in this paper Otsu's method is used too. Once an image is formed in grayscale form, next preprocessing step is applied on the image is the Binarization. Binary images are the images that have only values for each pixel, the two possible values are black and white. However, in this step a binary image is created from the original image to help to detect only the important areas and parts of Mushaf Al-Quran images. Fig. 12 illustrates the difference between normal image and binary image.

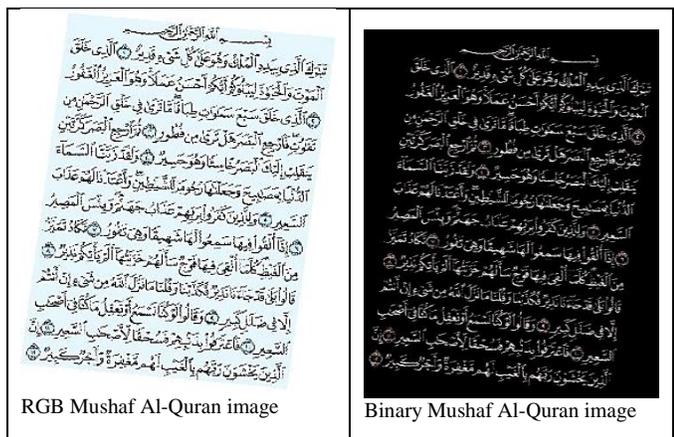


Fig. 12. Al-Quran Surah AlFajr (Different between color and binary image of Mushaf Al-Quran).

C. Foreground Mushaf Al-Quran Image Detection

Once a binary image is created, a morphology is applied to detect areas that have text in Mushaf Al-Quran images and then convert gotten text to lines using this morphology in the direction of x (close morphology is used here). This morphology produces image contains lines which will be used in the next stage for the angle calculation. Fig. 13 illustrates the difference between color image and foreground image.

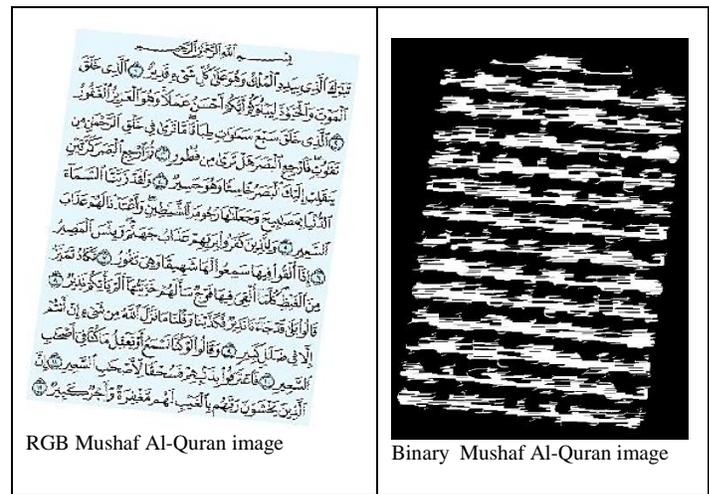


Fig. 13. Al-Quran Surah AlFajr (Color and Foreground Image Detection).

D. Angle Calculation

This is the most important stage where skew angle is calculated. Text in previous stage is converted to connected lines, so line detection comes second. The connected words in the previous stage can be considered as straight lines which helps to apply Hough transform method for line detection. To make it clear, this stage can be achieved in by two important steps as the following

1) *Line detection*: using (1) helps to detect straight lines in the images using the equation.

$$r = x \cos \theta + y \sin \theta \tag{1}$$

Hough transform is one of the most used feature extraction technique in computer vision, image analysis and digital image. It was introduced by Paul Hough 1962. So based on the Fig. 14 for each point (X₀, Y₀) there are other set of points which can create lines as (X₁, Y₁), this set of points that create line can be defined with equation (1). The two values (r, θ) in the above formula represents the lines (connected words) in Mushaf Al Quran images that gotten from the previous stage) that goes within (X₀, Y₀). In other words, the line in the image space represented as a point in the parameter space as in Fig. 15.

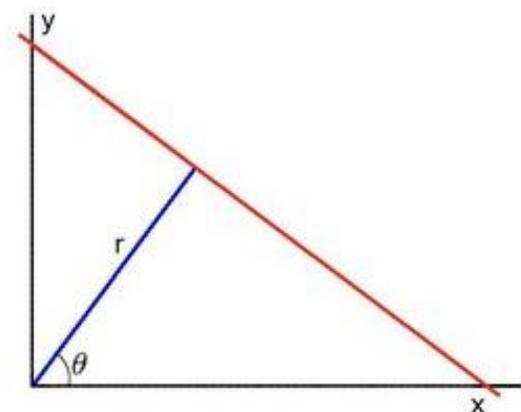


Fig. 14. Hough Transform Space.

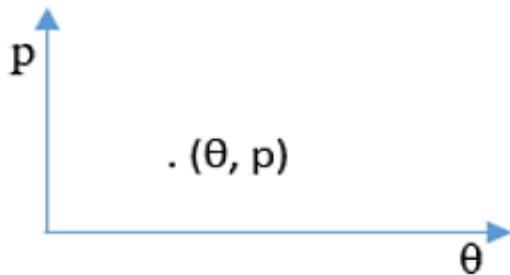


Fig. 15. Hough Transform Space.

Likewise, for linear Hough transform, two dimensional array are used for detecting lines in the image space in which each line is represented with two values of (θ, p) respectively. Further, straight lines are represented with peak strong point in the accumulator array in the image space as in Fig. 15 above. Once all peak points are detected, then it's easy to find line segments by end points of the peak values. The more intersections in the image space leads to the longest line among lines and that is the required line to calculate the skew angle.

2) *Angle Calculation*: skew angle is calculated using the longest straight line. In other words, it can be calculated with the deviation of the line with horizontal axis.

3) *Rotate image*: Once the skew angle is detected, it becomes easy for rotate it in order to correct the skew. However, several methods are used for skew correction like (direct method, indirect method contour oriented projection based and others), rotation of the image is done through Affine Transformation (2).

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (2)$$

Where (x, y) are the coordinates of the skew detected line, (θ) is the angle detected by Hough transform method. The above equation is for counter-clockwise.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(-\theta) & -\sin(\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (3)$$

Equation (2) in which the rotate the calculated skew angle to horizontal angle. The line is rotated with θ angle, if the detected angle is positive, the angle is corrected to the negative angle with the same value and the vice versa. Fig. 16 illustrates the last stage of the proposed method "Rotate image".

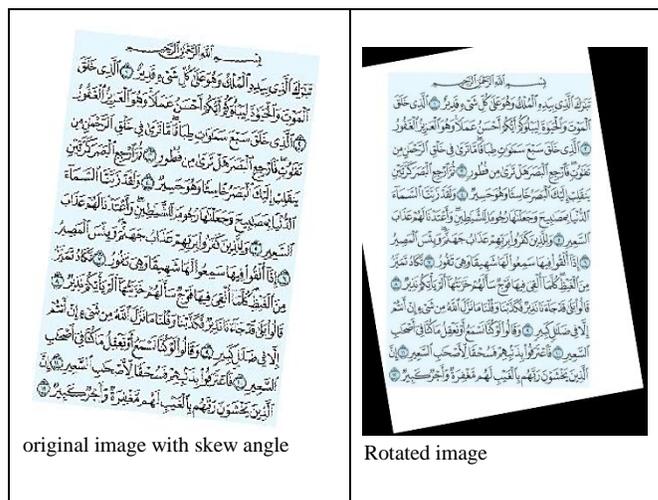


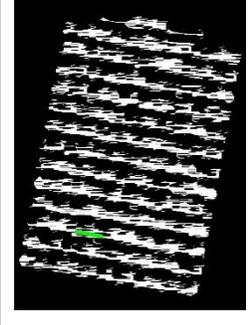
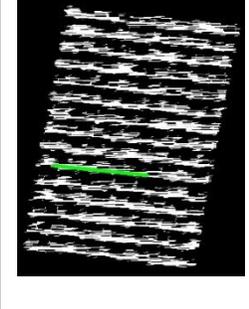
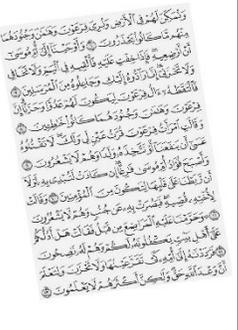
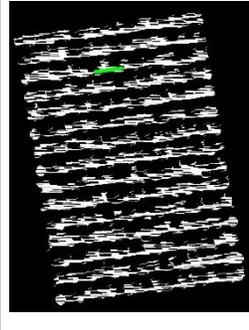
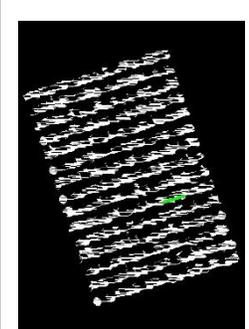
Fig. 16. Image Rotation.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this paper, a proposed method was tasted on Mushaf Al-Quran images that have skewed text. In addition, 50 Mushaf Al-Quran images were tasted by our proposed method. The proposed method has been implemented using Java programming language, Test environment used a PC with Intel i5- 2430M CPU @2.40 GHz with 12GB of memory, also Opencv function was implemented with Java code for detecting skew lines in the image as well as for measuring skew angle. In addition, the documents image of Mushaf Al-Quran were self-obtained from the source (<https://www.nourelquran.com/quranforall/fahd/index.php> [19]) and they were manipulated with different skew angle using software ImageJ. The accuracy for skew correction was about 90% for the images been tasted. Mostly, The Mushaf Al-Quran images that are colorful or have high resolution have lower accuracy in skew correction conversely with the images that have lower resolution in which the proposed method works perfectly. Therefore, Mushaf Al-Quran images have to be pre-processed before applying Hough transform method on, as in Fig. 11, converting input image to grayscale image helps to increase skew detection and process. Binary image also is a good way to increase skew detection as it was explained in proposed method section. The proposed method detecting and correcting skew angles through six stages namely, convert to grayscale, binary image, foreground image detection, HT transform method, calculate skew angle, rotate image. Table I shows a sample Mushaf Al-Quran images before and with skew correction at different angles.

TABLE I. SAMPLE RESULTS FOR SKEW DETECTION AND CORRECTION USING THE PROPOSED METHOD

Deg°	Original image	Grayscale image	Binary image	Line detection	Straighten image
3					
7					
-9					
17					
6					

10					
9					
-8					
-17					

VI. CONCLUSION AND FUTURE WORK

In this paper, a methodology for Mushaf Al-Quran skew detection and correction was presented. Moreover, the proposed method was based on Hough Transform algorithm which simply used by different handwritten script skew correction. This method is tasted on handwritten Mushaf Al-Quran images that have skew text in. Furthermore, the proposed method consists of six stages are combined together to deliver the final result. To conclude, this method can be improved for further research to be more accuracy. A possible

future work is to enhance the proposed technique with respect to processing time as well as to skew angle estimation.

ACKNOWLEDGMENT

The authors thank the Ministry of Education for funding this study through the following grants: FRGS/1/2017/ICT02/FTMK-CACT/F00345. Gratitude is also due to Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

- [1] I. Ahmad, 'A technique for skew detection of printed arabic documents', Proc. - 10th Int. Conf. Comput. Graph. Imaging, Vis. CGIV 2013, pp. 62–67, 2013.
- [2] M. S. Azmi, K. Omar, M. F. Nasrudin, B. Idrus, and K. Wan Mohd Ghazali, 'Digit recognition for Arabic/Jawi and Roman using features from triangle geometry', AIP Conf. Proc., vol. 1522, pp. 526–537, 2013.
- [3] A. Al-Khatatneh, S. A. Pitchay, and M. Al-Qudah, 'A Review of Skew Detection Techniques for Document', Proc. - UKSim-AMSS 17th Int. Conf. Comput. Model. Simulation, UKSim 2015, pp. 316–321, 2016.
- [4] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, T. Breuel, and A. Dengel, 'A Novel Skew Detection and Correction Approach for Scanned Documents', DAS. IAPR Int. Work. Doc. Anal. Syst. (DAS-12), April 11-14, Santorini, Greece, no. 4, pp. 1–2, 2016.
- [5] A. M. Al-Shatnawi, 'A skew detection and correction technique for Arabic script text-line based on subwords bounding', 2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014, pp. 324–328, 2014.
- [6] L. O'Gorman, 'The Document Spectrum for Page Layout Analysis', IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, 1993.
- [7] R. N. Verma and L. G. Malik, 'Review of illumination and skew correction techniques for scanned documents', Procedia Comput. Sci., vol. 45, no. C, pp. 322–327, 2015.
- [8] L. B. Mahanta and A. Deka, 'Skew and Slant Angles of Handwritten Signature', pp. 2030–2034, 2013.
- [9] T. A. Jundale and R. S. Hegadi, 'Skew Detection and Correction of Devanagari Script Using Hough Transform', Procedia Comput. Sci., vol. 45, pp. 305–311, 2015.
- [10] M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, 'Arabic calligraphy classification using triangle model for Digital Jawi Paleography analysis', Proc. 2011 11th Int. Conf. Hybrid Intell. Syst. HIS 2011, pp. 704–708, 2011.
- [11] C. Sun and D. Si, 'Skew and slant correction for document images using gradient direction', Proc. Fourth Int. Conf. Doc. Anal. Recognit., vol. 1, pp. 142–146, 1997.
- [12] G. Kapogiannopoulos and N. Kalouptsidis, 'A fast high precision algorithm for the estimation of skew angle using moments', Proceeding of IASTED, 2002.
- [13] S. C. Hinds, J. L. Fisher, and D. P. D. Amato, 'A DOCUMENT SKEW DETECTION METHOD USING RUN-LENGTH ENCODING AND THE HOUGH TRANSFORM', pp. 464–468, 1990.
- [14] M. S. Azmi, K. Omar, M. F. Nasrudin, A. K. Muda, and A. Abdullah, 'Arabic Calligraphy Identification for Digital Jawi Paleography using Triangle Blocks', Proc. 2011 11th Int. Conf. Hybrid Intell. Syst. HIS 2011, no. July, pp. 704–708, 2011.
- [15] M. Attia, 'Arabic Orthography vs . Arabic OCR', vol. 1, 2000.
- [16] A. A. Aburas and M. E. Gumah, 'Arabic Handwriting Recognition : Challenges and Solutions Electrical and Computer Engineering Dept International Islamic University Malaysia Department of Information Technology , University Technology PETRONAS 2 . Pervious related Research Work', 2008.
- [17] M. S. Azmi, K. Omar, M. F. Nasrudin, and A. K. Muda, 'Features Extraction of Arabic Calligraphy using extended Triangle Model for Digital Jawi Paleography Analysis', vol. 5, pp. 696–703, 2013.
- [18] F. Kleber, M. Diem, and R. Sablatnig, "Robust Skew Estimation of Handwritten and Printed Documents Based on Grayvalue Images." In Pattern Recognition (ICPR), 2014 22nd International Conference on, pp. 3020-3025. IEEE, 2014.
- [19] Nourelquran.com. (2018). مجمع نسخة | الملك سورة | القرآن نور موقع. الشريف المصحف لطباعة فهد الملك [online] Available at: <https://www.nourelquran.com/quranforall/fahd/sora/67.html> [Accessed 20 Aug. 2018].

Review of Information Security Policy based on Content Coverage and Online Presentation in Higher Education

Arash Ghazvini, Zarina Shukur, Zaihosnita Hood
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia,
43600 UKM, Bangi, Selangor, Malaysia

Abstract—Policies are high-level statements that are equal to organizational law and drive the decision-making process within the organization. Information security policy is not easy to develop unless organizations clearly identify the necessary steps required in the development process of an information security policy, particularly in institutions of higher education that largely utilize IT. An inappropriate development process or replication of security policy content from other organizations could fail in execution. The execution of a duplicated policy could fail to act in accordance with enforceable rules and regulations even though it is well developed. Hence, organizations need to develop appropriate policies in compliance with the organization regulatory requirements. This paper aims to reviews policies from selected universities with regards to ISO 27001:2013 minimum requirements as well as effective online presentation. The online presentation review covers the elements of aesthetics, navigation and content presentation. The information on the security policy document resides on the universities' website.

Keywords—Information security policy; policy development; higher education

I. INTRODUCTION

The aim of information security is to protect the organization's information assets from any unauthorized access, disclosure or breaches. To enforce an effective information security, organizations need to develop good management practices comprising policies and controls [35]. Technical solutions provide support to protect information assets. However, technical solution alone cannot eliminate the risks of information leakage, modification or breaches. As this may cause significant loss, information security is critical to the business operation of most organizations, especially government and public bodies as the financial and non-financial costs are much greater compared to other organizations [37]. Similarly, information leakage or breaches may cause great losses for a higher education institution that store a large amount of student information within the management system, administrative systems and student portals [35], [38]. For example, a university credibility and integrity can be damaged due to illicit grade changes and constant difficulties with registration or financial systems [21].

The importance of information security and confidentiality in universities has been discussed since 1975 [36]. Universities and colleges are being targeted for cyber-attacks

due to two main reasons. First, due to a large amount of computer power possess by universities and colleges. And second, due to the open access, they make available to the public. Universities' networking infrastructures are not only available to staff and students but are also available to other students, visitors, and researchers worldwide. While providing access to the public and promoting information sharing, there should be a balance to ensure the security of information assets [21].

Information security and protection against internal risks are focal concerns in many organizations. Technological solutions alone cannot guarantee data protection against various threats. Even though there are advanced technologies, human factor still remains as the major risk to the integrity of information systems security [17][24]. At this point, numerous security experts believe that implementation of security policy and enforcement are the most sensible approach to protect information systems security [15] and the key to an effective security control program [15][22]. 'Development process' [13][26] and 'contents' of the security policy are the two elements that mainly determine the effectiveness of security policy [8][19] [12].

Protection of organizations' information which is progressively stored, processed and disseminated is becoming more intricate and challenging. This is even more complex for knowledge-intensive organizations including universities as teaching and research activities are becoming more dependent on the availability, integrity, and accuracy of electronic information resources. This paper intends to study how to write general outlines and the structure of what a policy should contain, rather than the content of information security policies [7]. In addition, the online presentations of the policies are also reviewed based on a principle of good design.

II. ROLE AND SCOPE OF THE INFORMATION SECURITY POLICY

The literature shows that the information security policy is gradually becoming a significant corporate document to protect the availability, confidentiality, and integrity of organization information resources. More specifically, it is argued that the policy document should establish the mechanism for an organization to proactively manage

information security [14]. Hence, an effective information security policy should define individual responsibilities, outline authorized and unauthorized use of the system, create room for users to report any suspected or identified threats to the system, clarify penalties in case of violations, and specify methods for updating a policy [7].

One of the most significant roles of information security policy is to precisely specify user's rights and responsibilities and to successfully communicate it to all users, to ensure there is a mutual and coherent understanding of information security that is embraced by the organization [11]. This eliminates excuses for employees who fail to follow and execute security practices aligned with the organization's policy [23]. As a result, policy document must act as a catalyst of employees' belief and behavior with respect to information security, and by doing so, it becomes the foundation of effective security management [7].

The objective of information security is to protect organizations' information assets from unauthorized uses, breaches, and disclosure. As defined by ISO/IEC 27001:2013, information security refers to the preservation of confidentiality, integrity, and availability of information. The goal is providing access to only those authorized personnel who need the access, keeping the information accurate and complete and making sure the information is available to the authorized user when they need it.

Proper management practices containing policies and controls should be established to ensure the effectiveness of implementation and enforcement of information security policy. According to ISO/IEC 27002:2013, information security policy aims to provide management with guidance and support in accordance with corporate requirements and regulations when dealing with information security. Undoubtedly, information security policy plays an important role to ensure the organization's well-being by protecting the information assets. However, the development and implementation process of an effective information security is unclear [9].

Due to lack of guidance, policy developers often refer to developed policies by other organizations, available commercial sources, or public templates from the Internet. Thus, the policy document created from such sources will not provide proper guidance for information security to protect that individual organization. Moreover, the developed policy may not be applicable to the threats and risks that they are supposed to mitigate, and thus they will not resolve the security issues that a particular organization is facing. "Sadly, many IT security experts do not recognize and comprehend the business risks, and eventually make lengthy security policies documents that attempt to protect everything" [9].

The development process and implementing of an effective information security policy is not a clear cut and is triggered by various issues including regulatory requirements, complications of advanced technologies, internal and external risks and threats. The literature underlines a number of information security policy development process and implementation methods [1], although these methods do not

offer a comprehensive and integrated method that includes a step-by-step guideline [9].

III. INFORMATION POLICY STRUCTURE VS. POLICY GUIDELINE

Even though there is a substantial body of literature underlying the importance of the policy document, there is a debate on the structure and key elements of the policies. The literature has mostly explored the structure of policy, generally from a conceptual perspective. For instant reference [3] argue if there should be a single policy or whether it should be divided into subdocuments with different types. The previous study [29] proposes two models namely 'computer-oriented and people/organizational' policies. However, literature [30] suggests a three-level model that are 'institutional policy, institutional ISP and technical ISP'. In [31] recommends a four-level model including 'system security policy, product security policy, community security policy and corporate information security policy'. Whilst there is increasing debate about the number of policies and how they are inter-related, reference [31] state that practically organizations are more likely to have a single policy document. Other scholars are focusing on the difference between high and low levels of policy practices [32], although it should provide guidelines on 'means' as well as 'ends' [33]. Over the years, more studies have been conducted on the effective configuration for information security documentation, but surely minimum effort to resolve the issue. In fact, the issue has become even more complex due to the manifestation of new forms of security documents such as 'Internet and email usage policies' [2]; 'copyright policies' [18] that could complement the information security policy. As a result, there is a significant need for a focused, empirical study to examine the structural arrangements of information security policies, as they are currently being adapted and practiced by organizations [7].

The structure of information security policy has been largely discussed in the literature (although it lacks in empirical contributions and consensus). However, in academic, there is a fairly limited discussion about the particular issues that need to be addressed by the information security policy. The international standard 17799 ISO:2005 gives indications about the types of issues that can be addressed by information security policy, but the issues are less subjected to academic security. One of the very few attempts to precisely fill this gap was an empirical study by [7] about information security policies across large organizations in the UK, based on a framework where potential policy issues extracted from the literature. Even though the research offers useful insights, it lacks inconsistency of approach and terminology, because the study was drawn based on perceptions of IT decision makers about their own content of policy, rather than focusing on the actual content of policy [7].

In addition to concerns regarding the structure and content of policy, there are also concerns regarding policy effectiveness. Many organizations claim to have developed and implemented information security policy [20]. However, looking at the results, high degrees of information security

incidents and breaches suggest that there is a lack of effectiveness and/or communication of policy. In fact, the study by [34] revealed that there had been no significant changes in the number of security breached in organizations that had adopted an information security policy in comparison with those that had not. One possible reason for the ineffectiveness of information security policies is that organizations follow narrow policies that only focus on issues of information confidentiality, integrity, and availability. Unfortunately, infrastructure technology has failed to address increasingly important human and organizational aspects [6]. In fact, the most commonly adopted policy standard ISO 17799 (2005) @24) focus on the technically oriented conceptualization of information security (availability, confidentiality, and integrity), and ignores human factors such as trust, ethicality and the integrity of users [7].

A. Policy Writing Guidelines

Policies are high-level statements that correspond to corporate law that drives decision making in a university that is subject to a serious review process. The university's information security policies are accessible on their website. Standards are minimum requirements developed to address specific issues and requirements that ensure compliance with policies. Standards are used for verification purposes for audit and assessment. Every faculty and department are required to follow the standards and the adoption of local standards are encouraged to surpass the minimum requirements. A procedure is step-by-step instructions to accomplish certain tasks. Procedures can be also used to maintain compliance with regulations. Guidelines provide additional recommendations that provide a framework to help compliance with policies. They are more technical in nature compared to policies and standards. They are also updated more frequently to address changes in technology and university practices [28]. Fig. 1 presents the policy-making process.

Policy writing task should be done by reaching the intended audience with policies that are Clear, Easy to read and provide the right level of information to those affected by the content. If users understand a policy, they are more likely to follow it and incorporate it into their daily work. The key elements of a policy document are identified as 1) Policy Title, 2) Administrative Policy Statement Number and Functional Area, 3) Brief Description, 4) Applies To, 5) Reason for Policy, 6) Introduction, 7) Policy Statement, 8) Definitions, 9) Related Policies, Procedures, Forms, Guidelines, and Other Resources, 10) History, 11) Key Words [27].

- *Use Language That Reflects the Policy's Intent:*

Select the words carefully. Words like "should" and "may" imply a choice. For example, "Faculty and staff should not smoke in class." This means they shouldn't smoke but will be allowed if they do. The statement also does not address restrictions applicable to students. Examples of alternative phrasing would be: "Faculty, staff, and students are prohibited from smoking in class." this is much better, but only addresses a class setting. The best way to rewrite is "Smoking is not allowed inside University buildings".

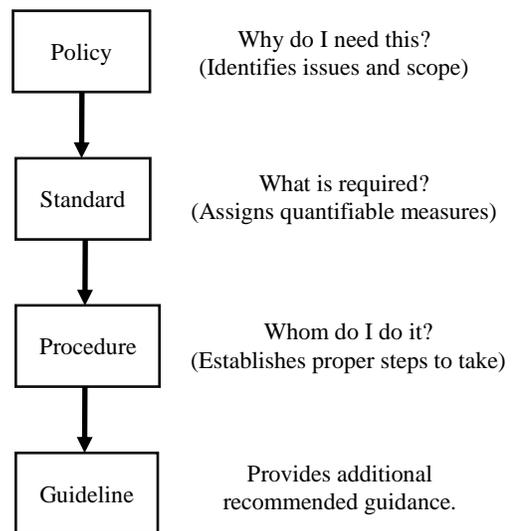


Fig. 1. Policymaking process

- *Use as Few Words as Possible to State a Case*

For instance, "All University faculty and staff, under the leadership of its officers, are obligated to ensure that University funds are used only for mission-related purposes." This statement implies that only those under the leadership are required to follow the policy. An alternative to the above statement is: "Employees must ensure that University funds are used only for mission-related purposes."

- *Ensure that Clarifying a Statement Did Not Alter Its Meaning:*

For example, "All faculty and staff must attend weekly meetings" The word "all" is redundant. Simply stating "Faculty and staff" implies all unless an exception is also written.

IV. REVIEW OF INFORMATION SECURITY POLICY DEVELOPMENT FRAMEWORKS

A. A Generic Framework for Information Security Policy Development

Reference [12] proposed a general framework to enhance security policies development process of higher education, using content analysis and cross-case analysis methods (Fig. 2). The proposed framework could be used as a guide to developing more comprehensive and sustainable information security policies in the institution of higher education. The framework can be used as a guideline to improve or develop a policy management program. However, the framework is too general, and it is necessary to explore more specific development processes such as the Acceptable Use Policy or any specific system security policy.

In [12] identified risk assessment as the major part policy development process since it systematically identifies, analyzes and evaluates the information security threats related to information systems and services as well as required controls to manage them. The process of risk identification involves identifying information assets, threats, and vulnerabilities. These are the important elements in identifying

the origin of incidents that could potentially affect the university information assets. The findings indicate that comprehension of security policy's content could be determined by the risk assessment.

B. The Policy Development Framework Including the ISPDLC Components

The result of a study by [9] shows that the most important of constructs is Risk Assessment (Fig. 3). Therefore, Risk Assessment should be the prior step in developing an

information security policy in order to identify the risks that need to be mitigated. Subsequently, Management Support is the second most important construct. Managers use policies to clarify their management intentions and direction. The result of the study also shows that Policy Monitoring was the least important construct. This suggests that the area of Policy Monitoring requires more attention. The content analysis implied similar results, with information security monitoring being the lowest frequency of tags among all categories.

Pre-Development	Development Process				Implementation	
Policy Team Development	Risk Analysis	Preparation	Writing Policy	Approval	Publish Strategy	Maintenance and Monitoring
<ul style="list-style-type: none"> Information Security Team Technical Writer Technical Personnel Legal Counsel Human Resources User Group (- Faculties, Centre, Department, Student representative, Vendor, Contractor) 	<ul style="list-style-type: none"> Identify internal & external threats Identify vulnerabilities Incidents/ Events Information asset <p style="text-align: center;">↓</p> <ul style="list-style-type: none"> Identify Issues 	<ul style="list-style-type: none"> Identify security control and legal requirement Identify characteristics of structure and cultural in organization Security practices Guidelines from security standards and best practices Benchmarking Create/ Review on existing policy 	<ul style="list-style-type: none"> Identify the policy contents and structure Draft language, style and formatting. Write initial draft Measure readability of policy document 	<ul style="list-style-type: none"> Review by additional stakeholder Obtain management endorsement & approval 	<ul style="list-style-type: none"> Plan communication Awareness program 	<ul style="list-style-type: none"> Plan maintenance Feedback Measure outcome Review and Update
Team	Risk Analysis	Preparation	Writing	Approval	Publish	Maintenance

Fig. 2. A generic framework for information security policy development.

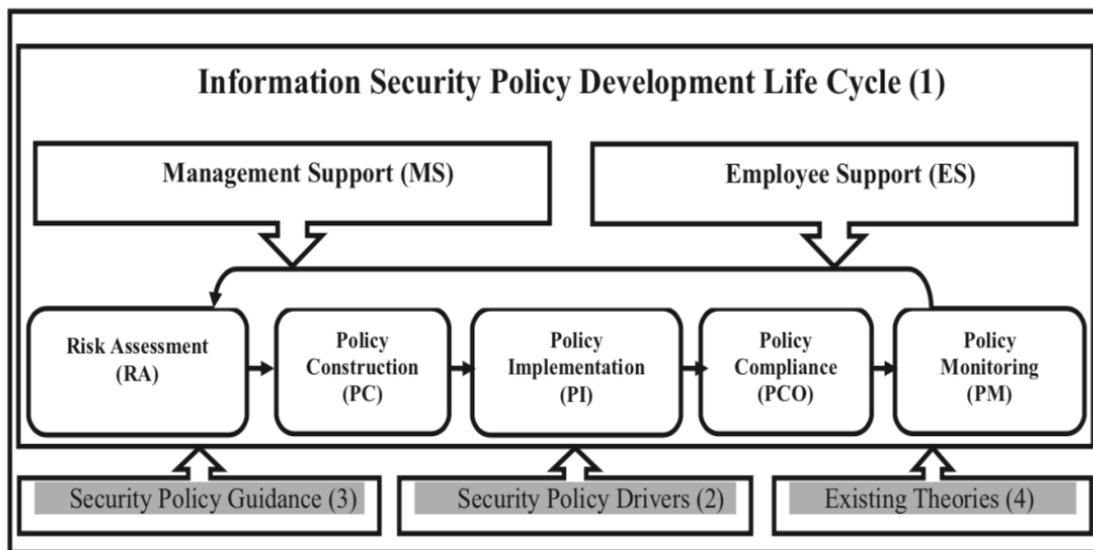


Fig. 3. The Policy development framework including the ISPDLC component.

The study by [9] has some limitations. The first one is the demographics of the respondents in the survey. The responded were only selected from the U.S. and the U.K. which makes it difficult to generalize the findings of the study, as the two countries are developed with advanced technology. Hence, while developing a framework, one should provide guidelines that can be adopted by both developed and underdeveloped countries to enhance their information security policy development process. In many developed countries, by law, senior managers or a board of directors are in charge of information security and risk management. Therefore, organizations have to spend resources to ensure the protection of an organization's information. However, this may not necessarily happen, especially in smaller organizations.

The second limitation is the time and cost involved in implementing the processes proposed in the framework. It requires organizations to have sufficient budget to cover all the costs such as the costs of conducting a risk assessment, constructing the information security policy, consulting with stakeholders, conducting training and education sessions and monitoring users' activities by, perhaps, using an automated monitoring system. Moreover, the costs are even higher for larger organizations as they require a significant amount of time and other resources. Lastly, the decision to develop and implement an information security policy should be based on organization security needs. Thus, a cost-benefit analysis should be carried on to understand whether it is worth for a particular organization to spend a large number of resources to do this exercise [9].

V. METHODOLOGY

As content analysis is helpful to identify trends and patterns in documents, this study focuses on two level of content analysis; first, to study information security policy development process for institutions of higher education, and second, to compare it to the common security information policy development adopted by organizations, which is discussed in the following sections. To fulfill this requirement, this study focused on the comparison of eleven universities' information security policy [12]. Information security policy is largely recognized as the most important information security mechanism to prevent, detect and respond to security

breaches. Therefore, it plays important role in IT-enable organizations especially defining the scope and content of information security policies. Each university's website was reviewed to identify the available policy documents and the information security coverage. Furthermore, the policies were reviewed in terms of aesthetics, navigation, and content.

A. University Selection

To ensure the consistency and accuracy of data collection from the information security policies of each university, a pro forma was devised. This pro forma was used to review the policies of eleven universities. The pro forma data collection document comprised the following four broad components:

- *University Details (Table I):*

Name, abbreviation, country, position in worldwide university ranking, website address; eleven universities have been selected from North America, Europe, Australia and Asia. All the selected universities are ranked below 250 worldwide, based on QS 2018 rankings.

- *Policy Administration Details (Table I):*

Details about the responsible department for the creation, management, and maintenance of the policy which includes responsible unit, phone number, and email address. Only responsible units are added to Table I to avoid invasion of personal privacy.

- *Policy structure (Table II):*

Types of available policy on the university website, besides the information security policy (e.g. Acceptable Use of Information Technology Resources Policy, Data Security Classification Policy).

- *Policy Coverage (Table II):*

Information security coverage and policy titles are listed here from each university's website. This task was cross-checked during the investigation by sending out emails to the respective university to ensure the accuracy and consistency. The contents of the pro forma were then summarized in Tables I and II to enable comparisons to be made.

TABLE I. UNIVERSITY AND POLICY ADMINISTRATION DETAILS

University	Abbrev.	University details			Responsible Unit
		Country	Ranking	Website	
University of Arizona	Arizona	United States of America	230	http://www.arizona.edu	UA Information Security
University of Minnesota	UMN	United States of America	163	https://twin-cities.umn.edu	UMN Office of Information Technology
Durham University	DUR	United Kingdom	78	https://www.dur.ac.uk	DUR IT Service Desk
University of Oxford	OX	United Kingdom	6	http://www.ox.ac.uk	OX University Council
University of Wollongong	UOW	Australia	232	https://www.uow.edu.au	UOW Information Management & Technology Services (IMTS)
Monash University	Monash	Australia	60	https://www.monash.edu	Monash IT Service Desk
University of Malaya	UM	Malaysia	114	https://www.um.edu.my	UM Information Technology Center
Universiti Kebangsaan Malaysia	UKM	Malaysia	230	http://www.ukm.my	UKM Information Technology Center

City University of Hong Kong	Cityu	Hong Kong	49	http://www.cityu.edu.hk	Cityu Information Security Unit
The Chinese University of Hong Kong	CHUK	Hong Kong	46	http://www.cuhk.edu.hk	CHUK Information Technology Services Center
National University of Singapore	NUS	Singapore	15	http://www.nus.edu.sg	NUS IT Care

B. Information Security Policies and Coverage

The introduction part of every university’ policy was helpful to understand its overall standpoint of information security. Some universities are concerned more about hardware protection or physical security, whereas other universities are more focused on confidentiality and integrity aspects of information assets and administrative data. There are some universities that emphasize the need for information for research. Therefore, they want to ensure security practices

help to promote research activities while protecting against attack. Because there are various areas of focus by different universities, we are not surprised to have found out there are also various policy structural arrangements and coverage. As illustrated in Table II the selected universities have different policies and the information security content coverage varies among them. The differences are determined during the risk analysis when the policy development team identifies the internal and external threats, vulnerabilities, incidents and information security assets.

TABLE II. POLICY TILES AND INFORMATION SECURITY COVERAGE

University	Policy Title	Information Security Coverage
University of Arizona	<ul style="list-style-type: none"> • General Information Security Policy • Computer and Network Access Agreement Policy • Acceptable Use of Computers and Networks Policy • Electronic Privacy Statement Policy 	<ul style="list-style-type: none"> • Information Security Policy • Asset Management • Human Resource Security • Physical and Environmental Security • Communications and Operations Management • Access Control • Information Systems Acquisition, Development, and Maintenance • Business Continuity Management • Compliance • Risk Assessment
University of Minnesota	<ul style="list-style-type: none"> • Acceptable Use of Information Technology Resources Policy • Data Security Classification Policy • Information Security Policy • Information Security Risk Management Policy • Internal Access to and Sharing University Information Policy • Reporting and Notifying Individuals of Information Security Breaches Policy • Including Privacy Statement on U Web Pages Policy 	<ul style="list-style-type: none"> • Acceptable Use of Information Technology Resources • Data Security Classification • Information Security • Information Security Risk Management • Internal Access to and Sharing University Information • Reporting and Notifying Individuals of Information Security Breaches • Including a Privacy Statement on U Web Pages
University of Durham	<ul style="list-style-type: none"> • Overarching Information Security Policies 1. Information Security Policy • Data Protection and Information Management Policies 1. Data Protection Policy 2. Records Management Policy and Records Retention Schedule • IT Regulations and Policies 	<ul style="list-style-type: none"> • Online Security • Data Handling • Responsibilities • Training and Advise
University of Oxford	<ul style="list-style-type: none"> • Data Protection: University Policy • Data Quality Policy • Freedom of Information Policy • Information Security Policy • Records Management Policy • Research Related Policy • Statement of Janet acceptable use policy 	<ul style="list-style-type: none"> • Access to the Janet for non-members • Advertising material on University web pages • Compliance • Disclaimer of liability • Disposal of old computers • Guidelines for handling illegal material • IT Rules • Mobile wireless networking regulations • Peer-to-peer resource sharing • Rules on mass mailing
University of Wollongong	<ul style="list-style-type: none"> • Cyber Security Policy • IT Acceptable Use Policy • IT Server Security Policy • Telephone and Mobile Use Policy 	<ul style="list-style-type: none"> • Computer Room Access • Cyber Security • IT Acceptable Use • IT User Account Management • Telephone and Mobile Use
Monash University	<ul style="list-style-type: none"> • Access to and Use of Electronic Resources Licensed by the Library Policy • Information Technology Acceptable Use Policy • Electronic Information Security Policy • Record-keeping Policy • Student Electronic Message Broadcast Policy • Web Accessibility Policy 	<ul style="list-style-type: none"> • Access to and Use of Electronic Resources Licensed by Library • Information Technology Acceptable Use • Electronic Information Security • ICT Security and Risk Management • Record-keeping • Student Electronic Message Broadcast • Web Accessibility

University	Policy Title	Information Security Coverage
university of Malay	<ul style="list-style-type: none"> • General Information Security Policy • ICT Security Policy • Wireless Communication Policy • Email Usage Policy • Server Colocation at PTM Data Centre Policy • Web Hosting Policy • Server handling Centre of Responsibility Policy • Firewall Policy • Malware Policy • Removal and Disposal of Media Policy • Supplier Management Policy • Source Code Management Policy • System Planning and Acceptance Policy • Termination Policy • Wireless Communication Policy 	<ul style="list-style-type: none"> • General • ICT Security • Network • Email • ICT Resources Management • Third Party / Vendor • Software • Website
Universiti Kebangsaan Malaysia	ICT Policies and Regulations	<ul style="list-style-type: none"> • Data Protection • Storage Security • Your Privacy • Information Collected • Policy Amendments
National University of Singapore	<ul style="list-style-type: none"> • IT Security Policy • Acceptable Use Policy 	<ul style="list-style-type: none"> • Information Security • Protect Your Computer • Protect Your Data • Protect your Privacy
City University of Hong Kong	<ul style="list-style-type: none"> • Policy on Use of IT Services and Facilities • Information Security Policy and Standards • Domain Name System Policy and Guidelines • Password Management Policy for User and System Accounts • Software Copyright Declaration and Compliance Observation 	<ul style="list-style-type: none"> • Use of IT Services and Facilities • Information Security and Standards • Domain Name System • Computer Account Retention for Leaving Staff • Retention for Deleted Email on MS Office 365 • Password Management for User and System Accounts
The Chinese University of Hong Kong	<ul style="list-style-type: none"> • University IT Policies • University IS Policies and Standards • Acceptable Use Policies and Guidelines 	<ul style="list-style-type: none"> • ICT Facilities & Services • OnePass Password Expiry • WiFi • Sharing Large Computer Equipment • Information Security • Display Name for Office 365 • Email Address for Staff • Computer Network, Access, and Usage • Email and the Internet Services • Data Centre and Networks • Computing Systems, Software and Account Information • Computer Laboratory

C. Online Presentation and Content Coverage

In [39] define aesthetic as the study of emotions and mind in the related notions such as the beautiful, the ugly as applicable to the fine arts. The aesthetic issue can influence user perception of a website. User's emotion and attitude can play an important role to attract the user's attention and keeping website trustworthy. Factor influencing the perception of beauty are balance proportion, informational content and complexity, contrast and clarity, and symmetry. Factors for aesthetic design features are visual complexity, color, and balance and symmetry [39].

In the case of navigation, it should lead the user to an easy, convenient and efficient browsing experience. Pagination navigation should not be invisible for users, hard to understand and difficult to identify [41]. In order to reduce the risk of users feeling disoriented and to assist them in finding information, navigation link should be the same from page to page [40].

The focus for content strategy is on the planning, creation, delivery, and governance content which might represent by text, images and multimedia [43]. Best practice for creating content meaningful identified by [43] are:

- Reflect your organization's goals and the user's needs.
- Understand how the user's think and speak about a subject.
- Communicate to people in a way that they understand.
- Be useful.
- Stay up-to-date and remain factual.
- Be accessible to all people.
- Be consistent.
- Be able to be found.
- Help define the requirements for the overall site.

In this study, the policies of 11 HEI Information Security Policies have been reviewed based on the criteria suggested by [42] as follows:

Aesthetics:

- What feel does the website give orderly or messy? Sparse or crowded? Playful or formal?
- Is the style consistent throughout the website?
- Where are photos or decorative touches getting in the way of my message?

Navigation:

- How easy is it to find information?
- Is there a search button for visitors?
- Do all the links work?

Content:

- Does the design make content easy to find?
- Will this content be relevant to the reader?
- Is the content concise but still useful?

TABLE III. UNIVERSITY WEBSITE AND CONTENT REVIEW

University	Aesthetics	Navigation	Content
University of Arizona	<ul style="list-style-type: none"> • Attractive and simple design – Orderly, sparse, formal. • The style is inconsistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Simple navigation without the need to guess • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is easy to find • Content is relevant • Content is concise but useful
University of Minnesota	<ul style="list-style-type: none"> • Appealing and simple design – Crowded but orderly, formal. • The style is consistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Simple navigation without the need to guess • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is easy to find • Content is relevant • Content is comprehensive
University of Durham	<ul style="list-style-type: none"> • Simple design – Orderly, sparse, formal. • The style is consistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Poor navigation - User can get lost in navigating between pages • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is not easy to find • Content is relevant but very brief in some cases • Content is presented in a form of: <ol style="list-style-type: none"> What do you know about this? What do you need to do? (Do..., Don't...) Where to next?
University of Oxford	<ul style="list-style-type: none"> • Attractive design – Orderly, sparse, playful. • The style is consistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Simple navigation without the need to guess • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is easy to find • Content is relevant • Content is comprehensive
University of Wollongong	<ul style="list-style-type: none"> • Attractive design – Orderly, sparse, playful. • The style is consistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Simple navigation without the need to guess • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is easy to find • Content is relevant • Content is concise but useful
Monash University	<ul style="list-style-type: none"> • Simple design – Orderly, Crowded, formal. • The style is inconsistent throughout the website • photos or decorative touches do 	<ul style="list-style-type: none"> • Poor navigation - User can get lost in navigating between pages as most links open in PDF • There is a search button only 	<ul style="list-style-type: none"> • Information is not easy to find – lack of good navigation and search button • Content is relevant • Content is concise but useful

University	Aesthetics	Navigation	Content
	not get in the way of the message	on the homepage • All links work	
University of Malay	<ul style="list-style-type: none"> • Appealing and simple design – Orderly, sparse, formal. • The style is consistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Simple navigation without the need to guess • There is a search button on the main page only • All links work 	<ul style="list-style-type: none"> • Information is not easy to find as the content is missing for some the policies and related documents • Content is relevant but not in single/default language. Some of the content is provided in English whereas the others in the Malay version. • Hyperlinks are not active for all PDF documents.
Universiti Kebangsaan Malaysia	<ul style="list-style-type: none"> • Appealing and simple design – Sparse and formal. • The style is consistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Poor navigation as information is spread across multiple pages without direct links • There is a search button • No links to connect the relevant pages • Some of the links do not work • Some link load PDF in the browser whereas the others download the PDF without permission 	<ul style="list-style-type: none"> • Information is not easy to find – Only covers UKM web security policy • Information security policies are presented as highlights and the content cannot be found • There is no default language as the English content is mixed with Malay version • Spelling mistakes – e.g. Guidelines • Does not state the objective and scope of UKM information security policy
National University of Singapore	<ul style="list-style-type: none"> • Appealing and simple design – Orderly, sparse, formal. • The style is consistent throughout the website • Photos or decorative touches can get in the way of the message 	<ul style="list-style-type: none"> • Poor navigation – Redundant and confusing navigation Panes • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is not easy to find – Only registered users are allowed to access the most of policies and guidelines. • Content is relevant but very brief in some cases • Content is presented in a form of: <ol style="list-style-type: none"> Protect Your Computer Protect Your Data Protect Your Privacy
City University of Hong Kong	<ul style="list-style-type: none"> • Simple design – Orderly, crowded, formal. • The style is inconsistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Simple navigation without the need to guess • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is easy to find • Content is relevant • Content is concise but useful
The Chinese University of Hong Kong	<ul style="list-style-type: none"> • Attractive design – Orderly, sparse, playful. • The style is consistent throughout the website • photos or decorative touches do not get in the way of the message 	<ul style="list-style-type: none"> • Simple navigation without the need to guess • There is a search button • All links work 	<ul style="list-style-type: none"> • Information is easy to find – Restricted access for some documents • Content is relevant • Content is comprehensive

Reviews from selected websites have been divided into three criteria aesthetics, navigation and content, as shown in Table III. Based on the table, we further highlight the existence of the respective criteria as shown in Table IV.

The strength of online presentation of this policies in terms of aesthetic elements are being attractive, orderly, sparse, simple, consistent, photos/decorative do not get in the way of the message, formal and appealing. However, some of the policies have issues in term of being inconsistent, crowded, playful and photos and decorative touches can get in the way of the message. Navigation strength of these policies are: simple navigation without the need to guess, search button available and link work.

Nonetheless, other identified issues are poor navigation where the user might get lost while searching for certain information, information is spread on multiple pages without a direct link, search functions are available on home page only, some link is not working and load pdf and download pdf without permission.

The strengths related to content are; easy to find, relevant content, concise but useful, and comprehensive. However, other identified issues are information not easy or cannot be found, brief and mixed, content is displayed in question and point form. Identified strengths from related websites can be a guide in order to design a good interface and avoiding some bad design issue of a website.

TABLE IV. ELEMENTS USED FOR AESTHETIC, NAVIGATION AND CONTENT CRITERIA

University	Aesthetic								Navigation			Content			
	Attractive	Orderly	Sparse	Simple	Consistent	Photos/ decorative do not get in the way of the message	Formal	Appealing	Simple Navigation	Search Button	Link Work	Easy to find	Relevant	Concise but useful	Comprehensive
University of Arizona	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	
University of Minnesota		✓		✓	✓	✓	✓	✓		✓	✓	✓	✓		✓
University of Durham		✓	✓	✓	✓	✓	✓		✓	✓	✓		✓		
University of Oxford	✓	✓	✓		✓	✓			✓	✓	✓	✓	✓		✓
University of Wollongong	✓	✓	✓		✓	✓				✓	✓	✓	✓	✓	
Monash University		✓		✓		✓	✓			✓	✓		✓	✓	
University of Malay		✓	✓	✓	✓	✓	✓	✓	✓	✓			✓		
Universiti Kebangsaan Malaysia			✓	✓	✓	✓	✓	✓		✓					
National University of Singapore		✓	✓	✓	✓		✓	✓		✓	✓		✓		
City University of Hong Kong		✓		✓		✓	✓		✓	✓	✓		✓	✓	
The Chinese University of Hong Kong	✓	✓	✓		✓	✓			✓	✓	✓	✓	✓		✓

Not all of 114 controls are mandatory as an organization can choose which controls are applicable and needs to be implemented and the rest could be declared as non-applicable. For example, the A.14.2.7 control, “Outsourced development” can be marked as non-applicable if the organization does not outsource any software development. The main criterion for selection of controls is the risk management as defined in clauses 6 and 8 of the ISO 27001.

ISO 27001:2013 Annex A is divided into three sections of mandatory documents, mandatory records and non-mandatory documents. Table V presents the structure of controls for the organization to be used to improve the security of information assets. (Please note that documents from Annex A are mandatory only if there are risks which would require their implementation).

TABLE V. ISO 27001:2013 ANNEX A MANDATORY AND NON-MANDATORY DOCUMENTS AND RECORDS

Mandatory documents required by ISO 27001:2013	Non-mandatory documents and records required by ISO 27001:2013	Mandatory records required by ISO 27001:2013
<ol style="list-style-type: none"> 1. The scope of the ISMS (clause 4.3) 2. Information security policy and objectives (clauses 5.2 and 6.2) 3. Risk assessment and risk treatment methodology (clause 6.1.2) 4. Statement of Applicability (clause 6.1.3 d) 5. Risk treatment plan (clauses 6.1.3 e and 6.2) 6. Risk assessment report (clause 8.2) 7. Definition of security roles and responsibilities (clauses A.7.1.2 and A.13.2.4) 8. Inventory of assets (clause A.8.1.1) 9. Acceptable use of assets (clause A.8.1.3) 10. Access control policy (clause A.9.1.1) 11. Operating procedures for IT management (clause A.12.1.1) 12. Secure system engineering principles (clause A.14.2.5) 13. Supplier security policy (clause A.15.1.1) 14. Incident management procedure (clause A.16.1.5) 15. Business continuity procedures (clause A.17.1.2) 16. Statutory, regulatory, and contractual requirements (clause A.18.1.1) 	<ol style="list-style-type: none"> 1. Procedure for document control (clause 7.5) 2. Controls for managing records (clause 7.5) 3. Procedure for internal audit (clause 9.2) 4. Procedure for corrective action (clause 10.1) 5. Bring your own device (BYOD) policy (clause A.6.2.1) 6. Mobile device and teleworking policy (clause A.6.2.1) 7. Information classification policy (clauses A.8.2.1, A.8.2.2, and A.8.2.3) 8. Password policy (clauses A.9.2.1, A.9.2.2, A.9.2.4, A.9.3.1, and A.9.4.3) 9. Disposal and destruction policy (clauses A.8.3.2 and A.11.2.7) 10. Procedures for working in secure areas (clause A.11.1.5) 11. Clear desk and clear screen policy (clause A.11.2.9) 12. Change management policy (clauses A.12.1.2 and A.14.2.4) 13. Backup policy (clause A.12.3.1) 14. Information transfer policy (clauses A.13.2.1, A.13.2.2, and A.13.2.3) 15. Business impact analysis (clause A.17.1.1) 16. Exercising and testing plan (clause A.17.1.3) 17. Maintenance and review plan (clause A.17.1.3) 18. Business continuity strategy (clause A.17.2.1) 	<ol style="list-style-type: none"> 1. Records of training, skills, experience, and qualifications (clause 7.2) 2. Monitoring and measurement results (clause 9.1) 3. Internal audit program (clause 9.2) 4. Results of internal audits (clause 9.2) 5. Results of the management review (clause 9.3) 6. Results of corrective actions (clause 10.1) 7. Logs of user activities, exceptions, and security events (clauses A.12.4.1 and A.12.4.3)

TABLE VI. MANDATORY DOCUMENTS REQUIRED BY ISO 27001:2013

Mandatory documents required by ISO 27001:2013	Arizona	UMN	DUR	OX	UOW	Monash	UM	UKM	NUS	Cityu	CHUK
The scope of the ISMS (clause 4.3)	X	X	X	X	X	X	X	X	X	X	X
Information security policy and objectives (clauses 5.2 and 6.2)	√	√	√	√	√	√	√	√	√	√	√
Risk assessment and risk treatment methodology (clause 6.1.2)	√	√	√	√	X	X	X	X	X	X	X
Statement of Applicability (clause 6.1.3 d)	√	X	X	X	X	X	X	X	X	X	X
Risk treatment plan (clauses 6.1.3 e and 6.2)	X	X	X	X	X	X	X	X	X	X	X
Risk assessment report (clause 8.2)	X	X	X	X	X	X	X	X	X	X	X
Definition of security roles and responsibilities (clauses A.7.1.2 and A.13.2.4)	X	X	X	√	X	√	X	X	X	X	X
Inventory of assets (clause A.8.1.1)	X	X	X	X	X	X	X	X	X	√	X
Acceptable use of assets (clause A.8.1.3)	√	√	X	X	√	√	X	X	√	√	X
Access control policy (clause A.9.1.1)	√	√	√	√	√	√	√	X	X	√	√
Operating procedures for IT management (clause A.12.1.1)	√	X	X	√	X	X	X	X	X	√	X
Secure system engineering principles (clause A.14.2.5)	X	X	X	X	X	X	X	X	X	X	X
Supplier security policy (clause A.15.1.1)	X	X	X	√	X	X	√	X	X	√	X
Incident management procedure (clause A.16.1.5)	√	X	X	√	X	X	X	X	X	√	X
Business continuity procedures (clause A.17.1.2)	√	X	X	X	X	X	X	X	X	√	X
Statutory, regulatory, and contractual requirements (clause A.18.1.1)	X	X	X	X	X	X	X	X	X	X	X
Total documents found out of 16 mandatory required documents	8	4	3	7	3	4	3	1	2	8	2

The selected universities' policies were reviewed in order to investigate the compliance with mandatory and non-mandatory documents and records by ISO 27001:2013. This task was cross-checked during the investigation by sending out emails to the respective university to ensure the accuracy and consistency. The findings were then summarised in Tables VI, VII and VIII to enable comparisons to be made. Table VI results show that none of the selected universities complied with all mandatory and no-mandatory documents and records from ISO 27001 Annex A.

This is again due to the policy development process, where the risk analysis task gives direction to policymakers to focus on certain information security issues. For instance, the University of Arizona made 8 out of 16 mandatory annex A documents available on the university's website, whereas the University Kebangsaan Malaysia has only 1 document available to be accessed by the visitors. Developing and dividing the information security content into standalone documents makes it easier to deliver the message to the intended audience and make the process more efficient.

TABLE VII. NON-MANDATORY DOCUMENTS AND RECORDS REQUIRED BY ISO 27001:2013

Non-mandatory documents and records required by ISO 27001:2013	Arizona	UMN	DUR	OX	UOW	Monash	UM	UKM	NUS	Cityu	CHUK
Procedure for document control (clause 7.5)	√	X	X	X	X	X	X	X	X	X	X
Controls for managing records (clause 7.5)	X	X	X	X	X	√	X	X	X	X	√
Procedure for internal audit (clause 9.2)	X	X	X	X	X	X	X	X	X	X	X
Procedure for corrective action (clause 10.1)	X	X	X	X	X	X	X	X	X	X	X
Bring your own device (BYOD) policy (clause A.6.2.1)	X	X	√	X	√	X	X	X	X	X	X
Mobile device and teleworking policy (clause A.6.2.1)	X	X	√	X	X	X	X	X	X	X	X
Information classification policy (clauses A.8.2.1, A.8.2.2, and A.8.2.3)	√	X	√	X	X	√	X	X	X	√	X
Password policy (clauses A.9.2.1, A.9.2.2, A.9.2.4, A.9.3.1, and A.9.4.3)	√	X	√	X	X	X	X	X	X	√	√
Disposal and destruction policy (clauses A.8.3.2 and A.11.2.7)	X	X	X	X	X	X	X	X	X	X	X
Procedures for working in secure areas (clause A.11.1.5)	X	X	X	X	X	X	X	X	X	X	X
Clear desk and clear screen policy (clause A.11.2.9)	X	X	X	X	X	X	X	X	X	X	X
Change management policy (clauses A.12.1.2 and A.14.2.4)	X	X	X	X	X	X	X	X	X	√	X
Backup policy (clause A.12.3.1)	X	X	√	X	X	X	X	X	X	√	X
Information transfer policy (clauses A.13.2.1, A.13.2.2, and A.13.2.3)	√	√	√	X	X	X	X	X	X	X	X
Business impact analysis (clause A.17.1.1)	√	X	X	X	X	X	X	X	X	X	X
Exercising and testing plan (clause A.17.1.3)	X	X	X	X	X	X	X	X	X	X	X
Maintenance and review plan (clause A.17.1.3)	√	X	√	X	X	X	X	X	X	√	√
Business continuity strategy (clause A.17.2.1)	√	X	X	X	X	X	X	X	X	√	X
Total documents found out of 18 mandatory required documents	7	1	7	0	1	2	0	0	0	6	3

TABLE VIII. MANDATORY RECORDS REQUIRED BY ISO 27001:2013

Mandatory records required by ISO 27001:2013	Arizona	UMN	DUR	OX	UOW	Monash	UM	UKM	NUS	Cityu	CHUK
Records of training, skills, experience, and qualifications (clause 7.2)	√	X	√	√	X	X	X	X	X	X	X
Monitoring and measurement results (clause 9.1)	X	X	√	X	X	X	X	X	X	X	X
Internal audit program (clause 9.2)	X	X	X	X	X	X	X	X	X	√	X
Results of internal audits (clause 9.2)	X	X	X	X	X	X	X	X	X	X	X
Results of the management review (clause 9.3)	X	X	X	X	X	X	X	X	X	X	X
Results of corrective actions (clause 10.1)	X	X	X	X	X	X	X	X	X	X	X
Logs of user activities, exceptions, and security events (clauses A.12.4.1 and A.12.4.3)	√	√	√	X	X	X	X	X	X	√	X
Total documents found out of 18 mandatory required records	2	1	3	1	0	0	0	0	0	2	0

VI. DISCUSSION

An effective information security policy should convert an organization's requirements into precise, measurable objectives that are readable and consistent [10]. Developing such information security policy that fulfills an organization's requirement is not easy an easy task. Duplicating a policy document from other organizations may not be sufficient to address issues such as compliance with regulatory requirements even though the replicated policy document is well-developed and properly referenced [16][3][4]. Thus, the security policy document must be developed based on the organization's culture, operations, environmental factors and policy requirement [25]. Therefore, the development process of information security policy should be tailored based on characteristics of the organizations, organizational culture, the potential technology changes in hardware and software, users and management support [5]. This applies to industries such as Higher Education where each university comprises diverse management structures, faculties, and departments, and practice different forms of behavior [21]. According to [13][9] studies often focus on the structure and content of policy but less on the development process, especially the step-by-step process. Hence, this paper exclusively focused on information security policy development in institutions of higher education [12].

If organizations seek to obtain ISO certification they must meet ISO 27001:2013 minimum requirement. These requirements are known as Annex A which includes mandatory and non-mandatory documents for organizations to create their policies based on. Many universities tend to develop a single document for all the policies and procedures (e.g. UKM), whereas other universities develop standalone policy documents based on ISO requirements. It is necessary to develop multiple policy documents because makes it possible to reach out to a targeted audience.

This paper conducted a comparative review of information security policy documents of eleven universities. The

objective is to review policy documents based on i) ISO 27001: 2013 mandatory and unmannerly requirements and ii) available frameworks and guidelines for the development of policy for higher education. The findings show that none of the selected universities have produced documents for all required mandatory and unmannerly requirements. This is due to risk analysis that should be the initial stage of policy development where the universities must identify the organization-specific issues as well as the organization regulatory agreements. Thus, developing a policy document for all Annex A requirements may not be necessary for every organization.

The information security policies must be accessible from the university website. However, not all policies should be accessible by the public. The policies should be divided into two categories including public and privet. The policies intended for the public must be accessible by everyone whereas the privet policies should be restricted by user authentication or require to be accessed within the university internal network. The privet policies are made for university stakeholders and internal use only. Making these policies accessible makes the organization vulnerable by giving an edge to those with prying eyes.

VII. CONCLUSION

The process of developing and implementing an effective information security policy is not a clear cut. It is vital for universities to realize the significance of the development process of information security policy for the institutions of higher education. The challenge for higher education institutions is to understand how to develop and implement information security policy effectively based on risk analysis in accordance with the organization's requirements. Otherwise, in case of security breaches or violations, it is less likely to enforce regulations due to incomplete or incomprehensible security policies document. This paper selected 11 universities to review their information security policies in contrast with ISO 27001:2013 minimum requirements to reach a concise understanding of the policy-

making process and what is being practiced in higher education. This study can be used as a guide for other universities who are developing or improving their information security policy to comply with ISO 27k series.

ACKNOWLEDGMENT

This study is supported by University Kebangsaan Malaysia (UKM). Grant code: AP-2017-003/2.

REFERENCES

- [1] V. Anand, J. Saniie, and E. Oruklu, "Security policy management process within six sigma framework," *J. Inf. Secur.*, vol. 3, no. 1, p. 49, 2012.
- [2] D. W. Arnesen and W. L. Weis, "Developing an effective company policy for employee internet and email use," *J. Organ. Cult. Commun. Confl.*, vol. 11, no. 2, p. 53, 2007.
- [3] R. Baskerville and M. Siponen, "An information security meta-policy for emergent organizations," *Logist. Inf. Manag.*, vol. 15, no. 5/6, pp. 337–346, 2002.
- [4] F. Bjorck, "Institutional theory: A new perspective for research into IS/IT security in organisations," in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, 2004, p. 5–pp.
- [5] S. C. Clark, R. A. Griffin, and C. K. Martin, "Alleviating the policy paradox through improved institutional policy systems: A case study," *Innov. High. Educ.*, vol. 37, no. 1, pp. 11–26, 2012.
- [6] G. Dhillon and G. Torkzadeh, "Value-focused assessment of information system security in organizations," *Inf. Syst. J.*, vol. 16, no. 3, pp. 293–314, 2006.
- [7] N. F. Doherty, L. Anastasakis, and H. Fulford, "The information security policy unpacked: A critical study of the content of university policies," *Int. J. Inf. Manage.*, vol. 29, no. 6, pp. 449–457, 2009.
- [8] N. F. Doherty, L. Anastasakis, and H. Fulford, "Reinforcing the security of corporate information resources: A critical review of the role of the acceptable use policy," *Int. J. Inf. Manage.*, vol. 31, no. 3, pp. 201–209, 2011.
- [9] S. V Flowerday and T. Tuyikeze, "Information security policy development and implementation: The what, how and who," *Comput. Secur.*, vol. 61, pp. 169–183, 2016.
- [10] S. Goel and I. N. Chengalur-Smith, "Metrics for characterizing the form of security policies," *J. Strategy. Inf. Syst.*, vol. 19, no. 4, pp. 281–295, 2010.
- [11] K.-S. Hong, Y.-P. Chi, L. R. Chao, and J.-H. Tang, "An empirical study of information security policy on information security elevation in Taiwan," *Inf. Manag. Comput. Secur.*, vol. 14, no. 2, pp. 104–115, 2006.
- [12] W. B. W. Ismail, S. Widyarto, R. A. T. R. Ahmad, and K. A. Ghani, "A generic framework for information security policy development," in *Electrical Engineering, Computer Science and Informatics (EECSI), 2017 4th International Conference on*, 2017, pp. 1–6.
- [13] N. B. L. Jr, "Information Security Policy Development: A Literature Review," *Int. J. Innov. Res. Inf. Secur.*, vol. 3, no. 04, pp. 1–7, 2016.
- [14] R. Saint-Germain, "Information Security Management Best Practice Based on ISO/IEC 17799," *Inf. Manag. J.*, vol. 39, no. 4, pp. 60–66, 2005.
- [15] K. J. Knapp, R. F. Morris Jr, T. E. Marshall, and T. A. Byrd, "Information security policy: An organizational-level process model," *Comput. Secur.*, vol. 28, no. 7, pp. 493–508, 2009.
- [16] R. P. Kusserow, "Developing and Managing Compliance Policy Documents," *J. Heal. Care Compliance—May–June*, p. 28, 2014.
- [17] B. Lebek, J. Uffen, M. Neumann, B. Hohler, and M. H. Breitner, "Information security awareness and behavior: a theory-based literature review," *Manag. Res. Rev.*, vol. 37, no. 12, pp. 1049–1092, 2014.
- [18] K. A. Loggie et al., "An analysis of copyright policies for distance learning materials at major research universities," *J. Interact. Online Learn.*, vol. 5, no. 3, pp. 224–242, 2006.
- [19] S. Maynard and A. B. Ruighaver, "What makes a good information security policy: a preliminary framework for evaluating security policy quality," in *Proceedings of the fifth annual security conference*, Las Vegas, Nevada USA, 2006, pp. 19–20.
- [20] J. Merete Hagen, E. Albrechtsen, and J. Hovden, "Implementation and effectiveness of organizational information security measures," *Inf. Manag. Comput. Secur.*, vol. 16, no. 4, pp. 377–397, 2008.
- [21] Y. Rezgui and A. Marks, "Information security awareness in higher education: An exploratory study," *Comput. Secur.*, vol. 27, no. 7–8, pp. 241–253, 2008.
- [22] N. S. Safa, R. Von Solms, and S. Furnell, "Information security policy compliance model in organizations," *Comput. Secur.*, vol. 56, pp. 70–82, 2016.
- [23] M. S. Saleh, A. Alrabiah, and S. H. Bakry, "Using ISO 17799: 2005 information security management: a STOPE view with six sigma approach," *Int. J. Netw. Manag.*, vol. 17, no. 1, pp. 85–97, 2007.
- [24] J. Shropshire, M. Warkentin, and S. Sharma, "Personality, attitudes, and intentions: Predicting initial adoption of information security behavior," *Comput. Secur.*, vol. 49, pp. 177–191, 2015.
- [25] M. Siponen and R. Willison, "Information security management standards: Problems and solutions," *Inf. Manag.*, vol. 46, no. 5, pp. 267–270, 2009.
- [26] T. Tuyikeze and S. Flowerday, "Information Security Policy Development and Implementation: A Content Analysis Approach," in *HAISA, 2014*, pp. 11–20.
- [27] University of Colorado, "User Guide to Writing Policies." [Online]. Available: <https://www.cu.edu/sites/default/files/APSwritingguide.pdf>.
- [28] "Policy and Guidance," University of Arizona. [Online]. Available: <https://security.arizona.edu/policy>.
- [29] M. T. Siponen, "Policies for construction of information systems' security guidelines," in *IFIP International Information Security Conference, 2000*, pp. 111–120.
- [30] D. F. Sterne, "On the Buzzword?? Security Policy??" 1991, p. 219.
- [31] K. R. Lindup, "A new model for information security policies," *Comput. Secur.*, vol. 14, no. 8, pp. 691–695, 1995.
- [32] B. Moule and L. Giavara, "Policies, procedures and standards: an approach for implementation," *Inf. Manag. Comput. Secur.*, vol. 3, no. 3, pp. 7–16, 1995.
- [33] J. Rees, S. Bandyopadhyay, and E. H. Spafford, "PFIREs: A Policy Framework for Information Security," *Commun. ACM*, vol. 46, no. 7, pp. 101–106, Jul. 2003.
- [34] N. F. Doherty and H. Fulford, "Do information security policies reduce the incidence of security breaches: an exploratory analysis," *Inf. Resour. Manag. J.*, vol. 18, no. 4, pp. 21–39, 2005.
- [35] S. K. S. Cheung, "Information Security Management for Higher Education Institutions," in *Intelligent Data analysis and its Applications, Volume I*, Springer, 2014, pp. 11–19.
- [36] B. Kerievsky, "Security and confidentiality in a university computer network," *ACM SIGUCCS Newsl.*, vol. 6, no. 3, pp. 9–11, 1976.
- [37] S. Singh and D. S. Karaulia, "E-governance: information security issues," in *Proceedings of the International Conference on Computer Science and Information Technology, 2011*, pp. 120–124.
- [38] H.-J. Kam, P. Katerattanakul, G. Gogolin, and S. Hong, "Information Security Policy Compliance in Higher Education: A Neo-Institutional Perspective.," in *PACIS, 2013*, p. 106.
- [39] J. Chen, "The Impact of Aesthetics on Attitudes Towards Websites," Sep-2013.
- [40] D. R. Danielson, "Transitional volatility in web navigation," *It Soc.*, vol. 1, no. 3, pp. 131–158, 2003.
- [41] M. Hu and Y. Kuang, "Human-machine interface: Design principles of pagination navigation in web applications," in *Computer Science & Education (ICCSE), 2014 9th International Conference on*, 2014, pp. 1140–1143.
- [42] S. Mallon, "5 Ways to Evaluate the Quality of Your Website Design.," straightnorth, 2018. [Online]. Available: <https://www.straightnorth.com/insights/5-ways-evaluate-quality-your-website-design/>.
- [43] Usability.gov, "Content Strategy Basics." [Online]. Available: <https://www.usability.gov/what-and-why/content-strategy.html>.

Implementation of a Formal Software Requirements Ambiguity Prevention Tool

Rasha Alomari

Computer Science Department
Faculty of Computing & Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Hanan Elazhary

Computer Science Department
Faculty of Computing & Information Technology
Jeddah University, Jeddah, Saudi Arabia
Computers and Systems Department
Electronics Research Institute, Cairo, Egypt

Abstract—The success of the software engineering process depends heavily on clear unambiguous software requirements. Ambiguity refers to the possibility to understand a requirement in more than one way. Unfortunately, ambiguity is an inherent property of the natural languages used to write the software user requirements. This could cause a final faulty system implementation, which is too expensive to correct. The basic requirements ambiguity resolution approaches in the literature are ambiguity detection, ambiguity avoidance, and ambiguity prevention. Ambiguity prevention is the least tackled approach because it requires designing formal languages and templates, which are hard to implement. The main goal of this paper is to provide full implementation of an ambiguity prevention tool and then study its effectiveness using real requirements. Towards this goal, we developed a set of Finite State Machine (FSMs) implementing templates of various requirement types. We then used Python to implement the ambiguity prevention tool based on those FSMs. We also collected a benchmark of 2460 real requirements and selected a random set of forty real requirements to test the effectiveness of the developed tool. The experiment showed that the implemented ambiguity prevention tool can prevent critical requirements ambiguity issues such as missing information or domain ambiguity. Nevertheless, there is a tradeoff between ambiguity prevention and the effort needed to write the requirements using the imposed templates.

Keywords—Software requirements; requirements ambiguity; natural language ambiguity; ambiguity prevention; controlled languages; finite state machines

I. INTRODUCTION

Software engineering passes through several subsequent stages. One of the preliminary stages is requirements elicitation from stakeholders. Unfortunately, elicited user requirements typically suffer from some imprecision challenging issues such as inaccuracy, inconsistency, incompleteness and ambiguity [1].

One of the most challenging issues is requirements ambiguity, which is an inherent characteristic of natural languages that are mostly used in writing software user requirements. Ambiguity occurs when an expression could have more than one way to be interpreted or understood. Consequently, it can lead to critical errors that pass through subsequent stages and end up with faulty software behavior [2, 3]. Paying attention to solving ambiguity problems in the

requirements elicitation stage is much easier and less expensive than correcting later software errors. For that, many research studies in the literature attempted to tackle this problem. There is no unified terminology in the literature for classifying techniques for ambiguity resolution. Accordingly, we adopt the following definitions:

- Ambiguity avoidance: denotes using rules and best practices while writing the requirements such as those proposed by Wiegers [3, 4].
- Ambiguity prevention: refers to forcing the users to write the requirements by filling in patterns or boilerplates corresponding to different types of requirements like the work of Stalhane and Wien [5] and Arora et al. [6].
- Ambiguity detection: refers to automatically detecting ambiguities after the user requirements are written like the work of Gleich et al. [7] and Wang et al. [8].
- Ambiguity correction: refers to semi-automated tools that interact with the user to make the needed corrections such as the work of Gill et al. [9].

One of the least tackled approaches is ambiguity prevention. A major drawback is that we could hardly find a fully implemented tool for this purpose, hindering its use in practice. Additionally, there is a shortage in empirical evaluations of such techniques [10]. The reason is that this approach requires developing and implementing formal representations. Hence, this is the main concern of the paper. The rest of the paper discusses related work in the literature. After that, the ambiguity prevention tool is detailed. Next, the experiment and discussion are provided; followed by the conclusion and future work.

II. RELATED WORK

In this section, we discuss some of the most prominent research studies in each of ambiguity avoidance, prevention, detection, and correction.

A. Ambiguity Avoidance

In ambiguity avoidance studies, the main methods used are rules and best practices. In this direction, Wiegers [3, 4] provided rules to avoid ambiguity, such as mentioning some

ambiguous words and expressions that should be avoided. Jain et al. [11] proposed a tool that can be implicitly considered an avoidance tool since it enforces requirements documentation best practices such as using standardized syntaxes and the consistent use of terminology; though it mainly falls into the ambiguity prevention class as discussed below.

B. Ambiguity Prevention

As previously noted, ambiguity prevention efforts use controlled natural languages such as templates, patterns, and boilerplates. For example, Jain et al [11] proposed a Requirements Analysis Tool (RAT) that uses templates to enforce requirements documentation best practices. RAT is comprised of a set of Finite State Machines (FSMs). It classifies the requirements into several types and then verifies that the requirements follow one of the best practice syntaxes supported by the tool. It then produces warning messages explaining where requirements are ambiguous and displays suggestions to fix them. This tool has been adopted in [12] for the Arabic language, and its full implementation is the main concern of this paper.

Denger et al. [13], on the other hand, proposed natural language patterns to be used by requirements authors when writing embedded systems requirements to prevent ambiguity. Farfeleder et al. [14] presented a tool that uses ontology-based reasoning to guide the requirements engineers and enforced this guidance by using boilerplates.

C. Ambiguity Detection

Gleich et al. [7] proposed a tool to automate the ambiguity detection process and explain the sources of detected ambiguities. It considers lexical, syntactic, semantic, and pragmatic ambiguity in addition to vagueness and language errors. This work uses part of speech tagging and regular expression search techniques for ambiguity detection. Similarly, the work of Wang et al. [15] automated the lexical ambiguity detection process focusing on overloaded and synonymous lexical ambiguity sources. The detection procedure goes through two main steps. In the first step, the C-value statistical method is used for terms extractions [16]. In the second step, the extracted terms are ranked according to the ambiguity score. The authors proposed features-based methods to estimate ambiguity scores. The ranking aims to help the requirements engineer to decide which ambiguities are more serious for time saving. Yang et al. [17] focused on one type of ambiguity, which is anaphoric ambiguity. Anaphoric ambiguity occurs when a linguistic expression may refer to two or more antecedent candidates. In this work, the authors introduced an architecture of an automated system to determine nocuous ambiguity and help requirements analysts to resolve it while discarding innocuous ambiguity that is unlikely to be misunderstood. Their approach relied on collecting human interpretations of instances of ambiguity, using heuristics to model human interpretations, and using machine learning to train the heuristics.

D. Ambiguity Correction

An example of ambiguity correction is the work of Gill et al. [9], who proposed a framework to develop semi-automatic tools for ambiguity correction in open source software

requirements. They discussed some challenges in open source requirements that make it a special case.

III. AMBIGUITY PREVENTION TOOL

As previously noted, ambiguity prevention approach uses controlled natural language such as templates, patterns, and boilerplates to prevent as much as possible ambiguity sources. We adopt the approach of Jain et al. [11], who uses templates, glossaries, and FSMs for this purpose. Templates are defined for six requirement classes. For each template, there is a matching FSM to analyze each requirement syntactically. Nevertheless, the authors provided merely details of the implementation of one requirements type. Hence practical use and adoption of the tool was hindered. We provide details of the implementation of all the FSMs.

In the following subsections, we explain the different requirement classes, the templates, the FSMs, the glossaries, and how the tool processes an input requirement through lexical analysis and syntactic analysis phases.

A. Requirements Classes

According to academic researchers and field experts, requirements can be classified into six classes. The six classes and their proposed syntaxes are shown and discussed below.

1) *Solution requirements*: This type of requirements expresses what an intended system or subsystem must do; for example:

Req01: *The system shall display completed work list items to the lab manager.*

2) *Enablement requirements*: Enablement requirements state what capabilities a proposed system or subsystems must provide to the users. There are two subcategories of enablement requirements. The first subcategory includes requirements that show an ability that should be provided by the software but does not decide which subsystem will provide it to the user. This is used when it is early to specify an exact ability provider; for example:

Req02: *Lab manager shall be able to create work list items.*

The second subcategory, on the other hand, includes more detailed requirements that state which system or subsystem should provide an ability to the user; for example:

Req03: *The system shall allow the lab manager to display work list items assigned to him, based on ID.*

3) *Action constraint requirements*: Those requirements define how the proposed system or subsystem is expected to act. There are two subcategories of action constraint requirements. The first subcategory includes requirements that state that the proposed system or some of its subsystems are allowed or not to do some action; for example:

Req04: *The loan subsystem may only delete a lender if there are no loans in the portfolio associated with this lender.*

The second subcategory, on the other hand, includes requirements that state business rules regarding how agents take some specific actions; for example:

Req05: Only library staff may perform the loan transactions.

4) *Attribute constraint requirements:* This requirements type is used to express constraints on an entity attributes or attribute values; for example:

Req06: Search options must always be one of the following: Price, Destination, Restaurant type, and Specific dish.

5) *Definition requirements:* This category is suitable to define entities as needed; for example:

Req07: The expected profit of a fixed rate loan is defined as the amount of interest received over the remaining life of the loan.

6) *Policy requirements:* This requirements type is used to illustrate the policies that must be followed by the system; for example:

Req08: Loan is not computed in more than one bundle.

B. Templates and Finite State Machines

Each requirements type has a specific template in addition to a corresponding FSM to determine whether an input token stream follows the syntax. We describe the FSMs using the following variables:

- Q denotes the set of states of a given FSM based on the syntax.
- S_0 is the start state, which is the same for all FSMs.
- F is the set of final states indicating that the input token stream was based on one of the syntaxes.
- E is the set of error states indicating that the input token stream did not follow any of the syntaxes.
- S denotes the alphabet set. It includes a set of modal phrases and keywords that differentiate the various FSMs. It also includes phrases from the entity and action glossaries described below. It is the same for all FSMs.
- δ is the transition function.

1) *Solution requirements FSM:* The solution requirements have one accepted template as follows; its FSM is depicted in Fig. 1:

<Agent Phrase> <"shall" | "must" | "will"> <Action Phrase>

2) *Enablement Requirements FSMs:* Enablement requirements have two accepted templates and therefore two FSMs. The first template is as follows:

<Agent Phrase> <"shall" | "must" | "will"> <"be able to"> <Action Phrase>

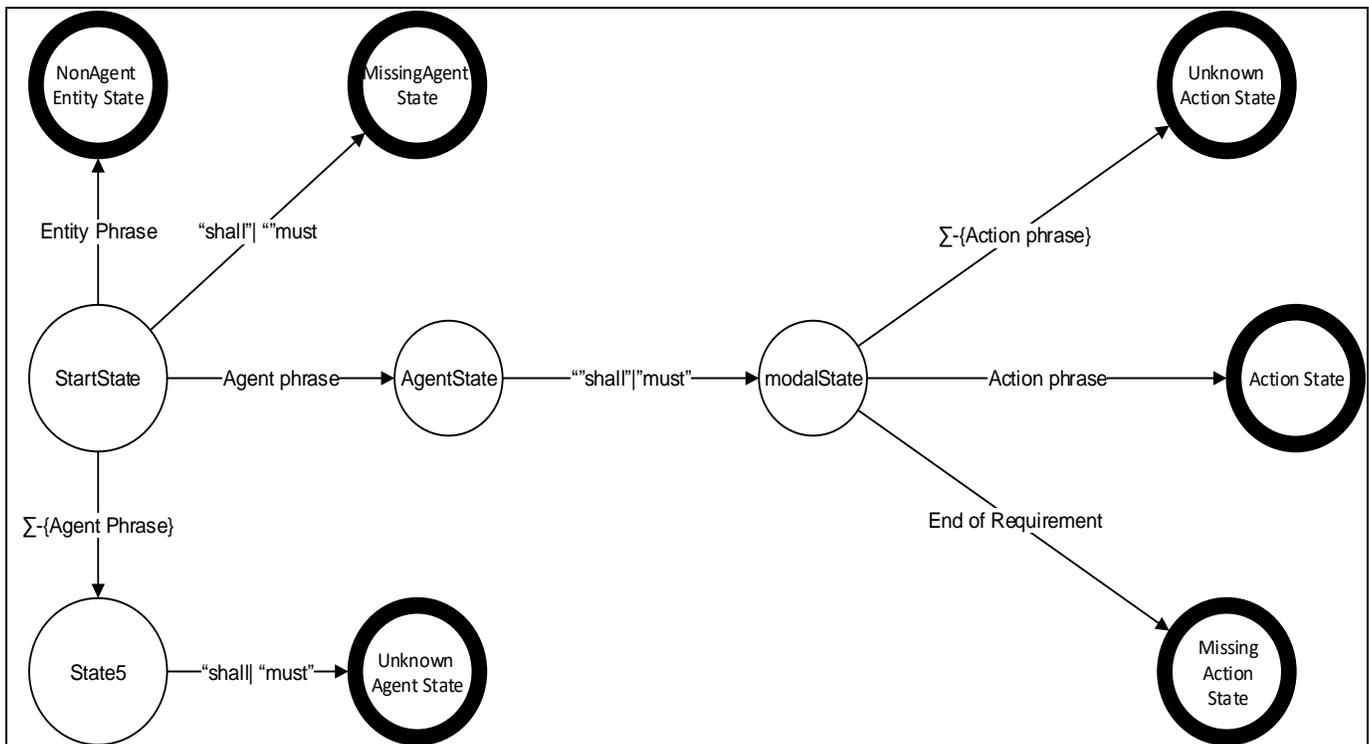


Fig. 1. Solution Requirements FSM.

The corresponding FSM is depicted in Fig. 2. It can be described as follows:

- $Q = \{\text{Start State, Action State, Modal State, Agent State, Missing Agent State, Missing Action State, Unknown Action State, Unknown Agent State, Non-Agent Entity State}\}$
- $F = \{\text{Action State}\}$
- $E = \{\text{Non-Agent Entity State, Missing Agent State, Unknown Action State, Missing Action State, Unknown Agent State}\}$

<Agent Phrase> <"shall" "must"|"will"> <"allow" | "permit"> <Agent Phrase><"to"> <Action Phrase>

The second accepted template of enablement requirements is as follows:

The corresponding FSM is depicted in Fig. 3. It can be described as follows:

- $Q = \{\text{Start State, Action State, Modal State, Agent State, Missing Agent State, Missing Action State, Unknown Action State, Unknown Agent State, Non-Agent Entity State}\}$
- $E = \{\text{Non-Agent Entity State, Missing Agent State, Unknown Action State, Missing Action State, Unknown Agent State}\}$
- $F = \{\text{Action State}\}$

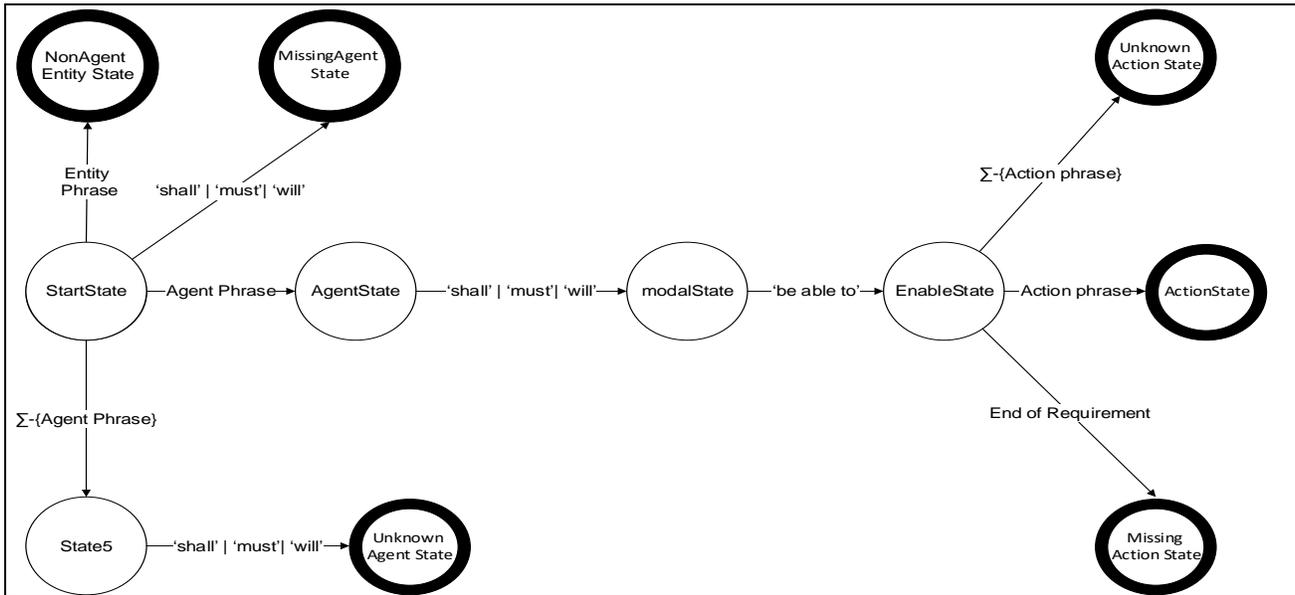


Fig. 2. Enablement Requirements FSM (1).

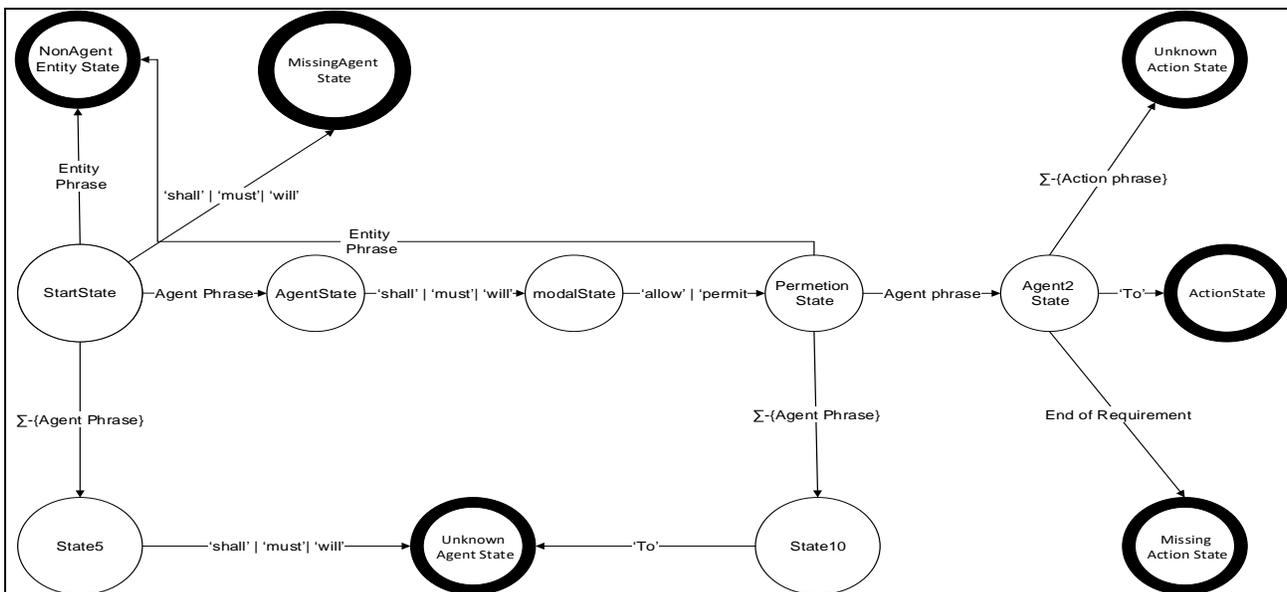


Fig. 3. Enablement Requirements FSM (2).

3) *Action constraint requirements FSMs*: Action constraint requirements have two accepted templates. The first one is as follows:

<Agent Phrase> <"shall" | "will" | "may"> <"only" | "not">
 <Action Phrase> <"when" | "if"> <condition>

The corresponding FSM is depicted in Fig. 4. It can be described as follows:

- Q = {Start State, Action State, Modal State, Agent State, Missing Agent State, Missing Action State, Unknown Action State, Unknown Agent State, Non-Agent Entity State}
- F = {Action State}

- E = {Non-Agent Entity State, Missing Agent State, Unknown Action State, Missing Action State, Unknown Agent State}

The second accepted template of action constraint requirements is as follows; the corresponding FSM is shown in Fig. 5:

"Only" <Agent Phrase> <"may" | "may be"> <Action Phrase>

4) *Attribute Constraint Requirements FSM*: Attribute constraint requirements have one accepted template:

<Entity Phrase | Agent Phrase> "must" <"always" | "never" | "not"> <"be" | "have"> <Value Phrase>

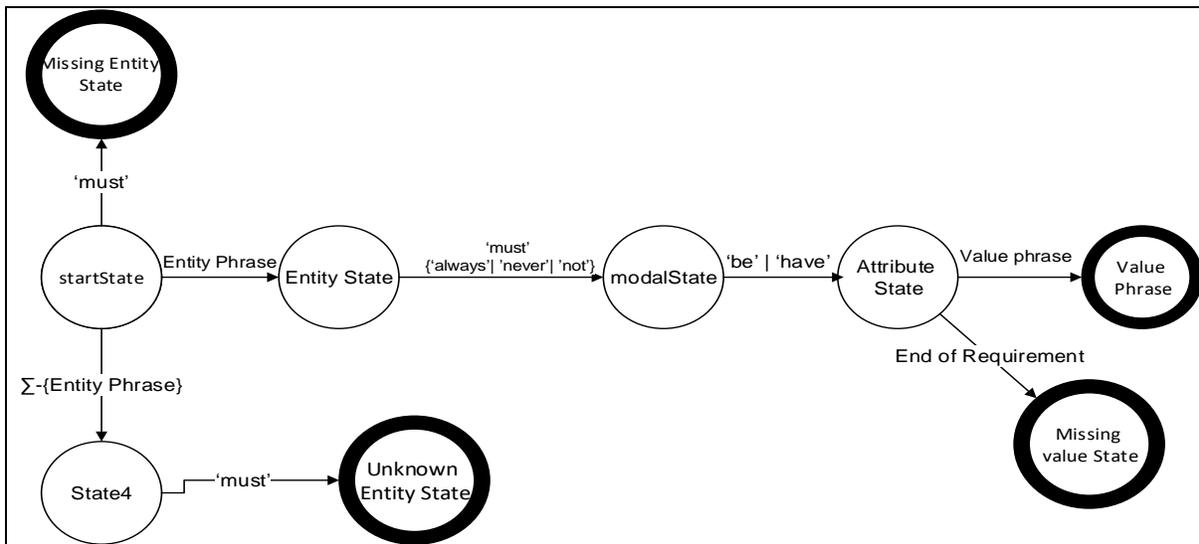


Fig. 4. Action Constraint Requirements FSM (1).

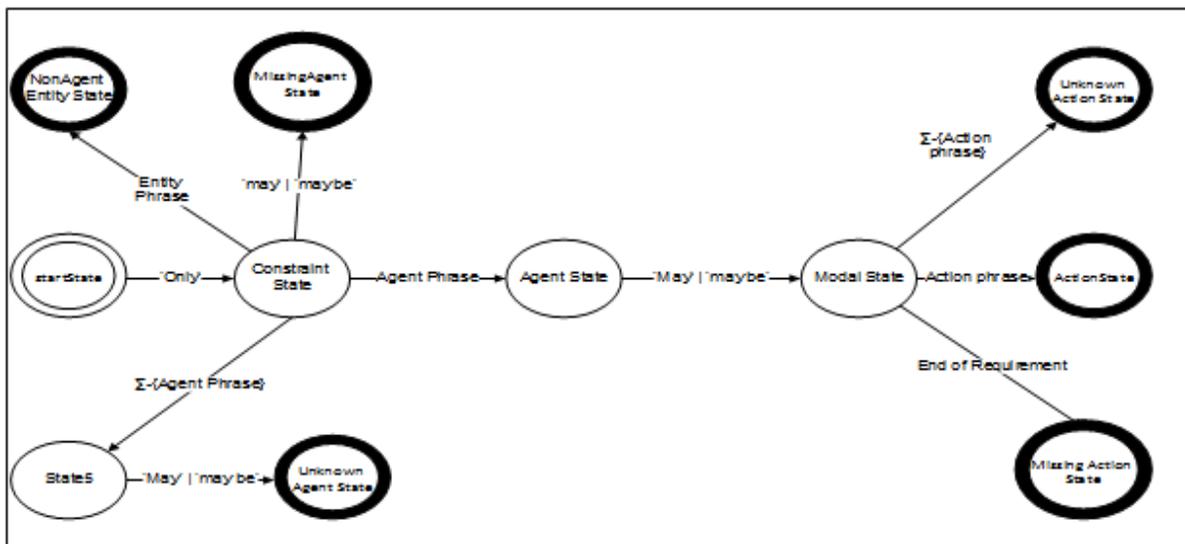


Fig. 5. Action Constraint Requirements FSM (2).

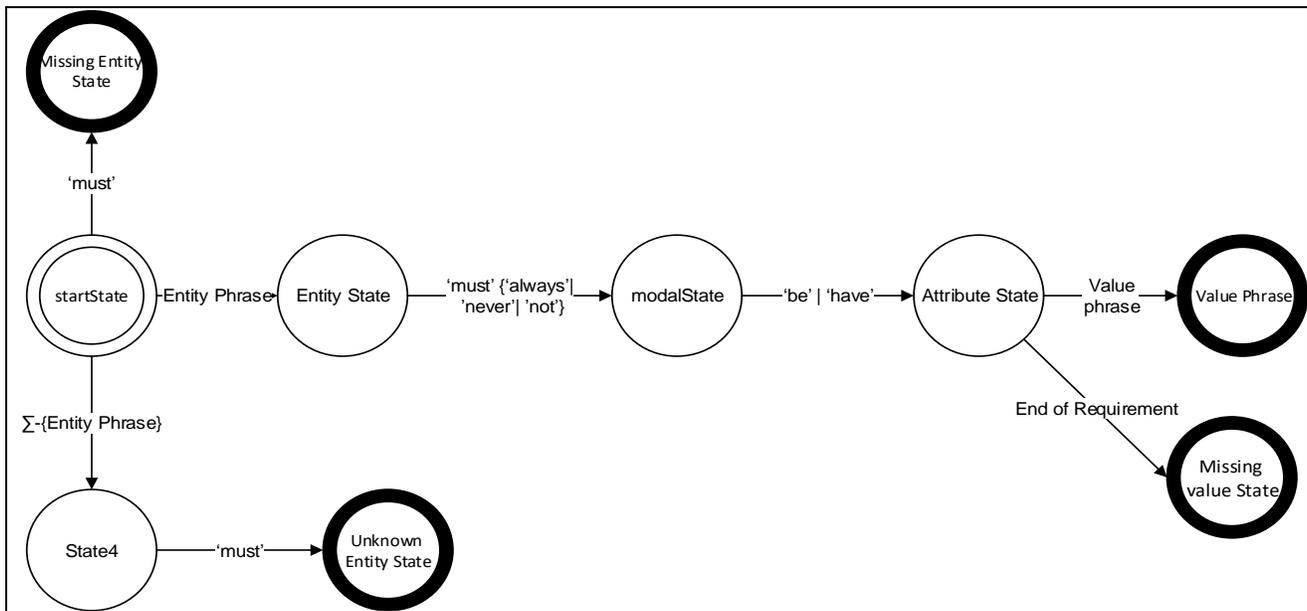


Fig. 6. Attribute Constraint Requirements FSM.

The corresponding FSM is depicted in Fig. 6. It can be described as follows:

- Q = {Start State, Action State, Modal State, Agent State, Missing Agent State, Missing Action State, Unknown Action State, Unknown Agent State, Non-Agent Entity State}
- F = {Action State}
- E = {Non-Agent Entity State, Missing Agent State, Unknown Action State, Missing Action State, Unknown Agent State}

5) *Definition requirements FSM*: Definition requirements have one accepted template as follows:

<Entity Phrase | Agent Phrase> <"is" | "will be"> <"defined as" | "classified as"> <Entity Phrase>

The corresponding FSM is depicted in Fig. 7. It can be described as follows:

- Q = {Start State, Action State, Modal State, Agent State, Missing Agent State, Missing Action State, Unknown Action State, Unknown Agent State, Non-Agent Entity State}
- F = {Action State}.
- E = {Non-Agent Entity State, Missing Agent State, Unknown Action State, Missing Action State, Unknown Agent State}

6) *Policy requirements FSM*: Policy requirements have one accepted template as follows:

<Entity Phrase | Agent Phrase> <"is" | "is not"> <Action Phrase>

The corresponding FSM is depicted in Fig. 8. It can be described as follows:

- Q = {Start State, Action State, Modal State, Agent State, Missing Agent State, Missing Action State, Unknown Action State, Unknown Agent State, Non-Agent Entity State}
- F = {Action State}.
- E = {Non-Agent Entity State, Missing Agent State, Unknown Action State, Missing Action State, Unknown Agent State}.

C. The Glossaries

The glossaries are an essential component of the implemented tool. The program consults user-defined glossaries to determine whether an input requirement uses predefined accepted terminology or not. Moreover, glossaries are necessary for lexical and syntactic analysis as described below.

We use two glossaries: an entity glossary and an action glossary. The entity glossary contains an entry for each accepted entity in the requirements document. Table I shows an example of an entity glossary content. The glossary determines whether each entity is an agent or not. An agent entity is the one that can do an action such as 'booking user' or 'library stuff', while a non-agent entity is an entity that does not perform an action such as 'the loan'. An action glossary, on the other hand, contains an entry for every accepted action phrase. Table II shows an example content of an action glossary.

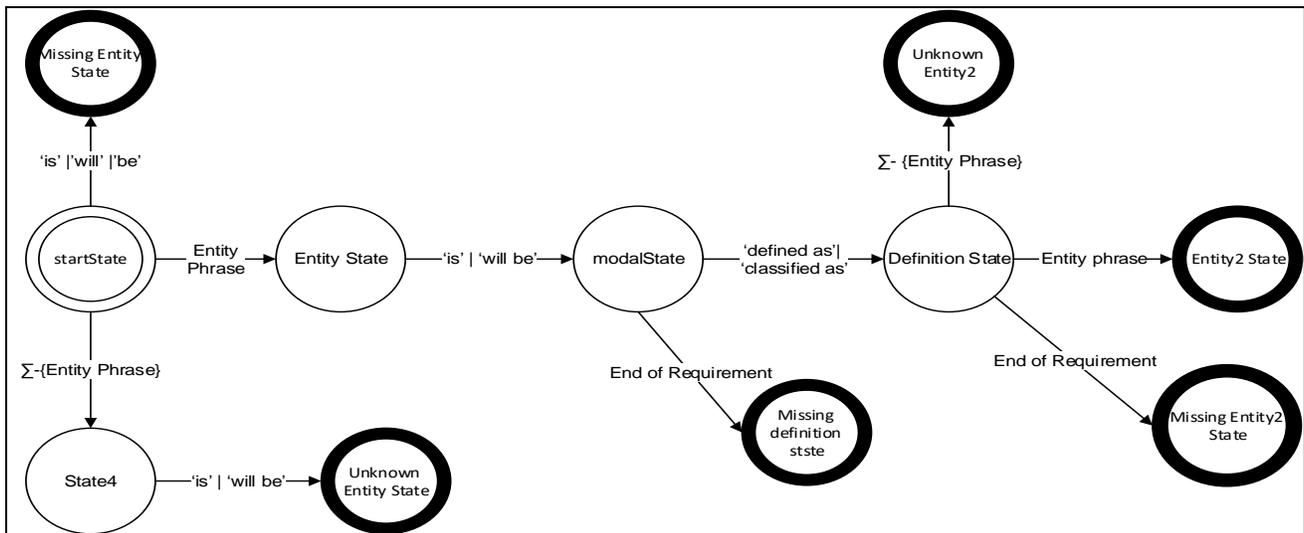


Fig. 7. Definition Requirements FSM.

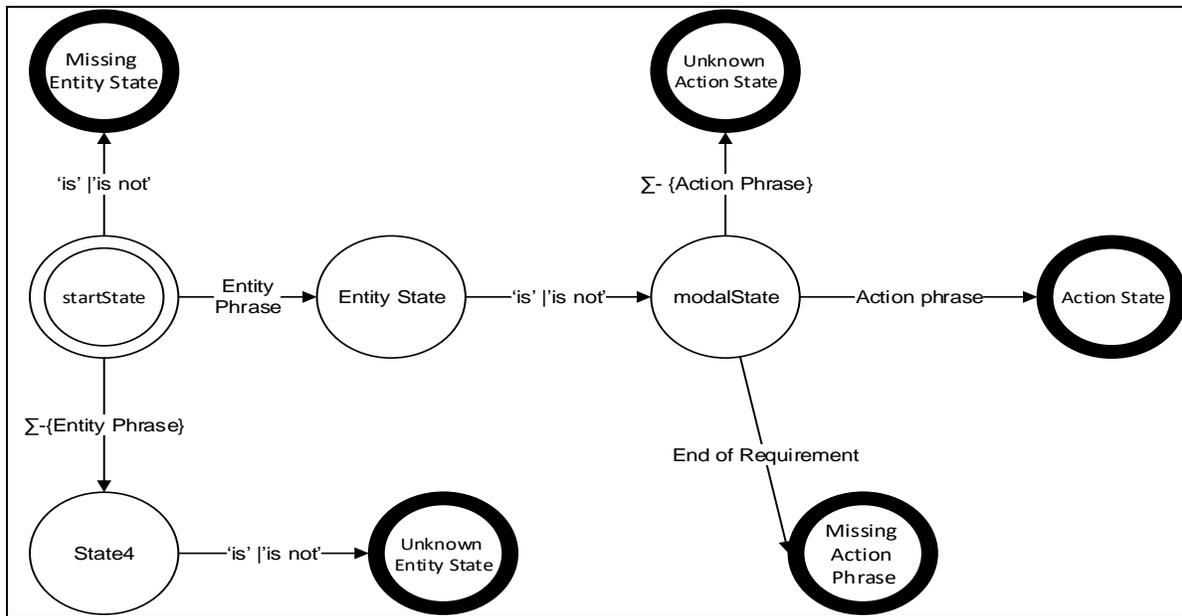


Fig. 8. Policy Requirements FSM.

TABLE I. EXAMPLE OF ENTITY GLOSSARY CONTENT

Entity Descriptor	Explanation	Is Agent
Borrower	The recipient of money from a lender. Borrowers may receive loans jointly; that is, each loan may have multiple borrowers.	Yes
HR User	User from human resource department	Yes
Protocol	the exact methodology used to analyze samples	No
ProdID	Product Identification; unique identifier of each product	No
Product Sample	A small amount of product taken from a specific product	No

TABLE II. EXAMPLE OF ACTION GLOSSARY CONTENT

Action Descriptor	Explanation
process orders	Action for processing orders
Display	Rendering an item on screen
send contracts data	Action for transfer of contract data
inform administrator	Action for sending e-mail notification to administrator
process payroll	Action for processing of payroll

TABLE III. TOKEN TYPES AND TAGS

Token Type	Tag
Label	Lbl
Entity phrase	En
Agent phrase	Ag
Action phrase	Ac
Modal phrase	Mod
Constant phrase	Const
Unknown	Un

D. Lexical Analysis

In the lexical analysis phase, the program consults the glossaries to tokenize a given requirement statement and then classify and tag each token into “entity phrase”, “agent phrase”, “action phrase”, “modal phrase”, “constant phrase”, or “unknowns” as depicted in Table III. The term constant phrase indicates phrases that do not fall into any of the other token types such as “be able to”, “only”, and “permit”. Stop words such as “the”, “a”, “an”, “for”, “too” and “up” are ignored in the process.

As an example, to clarify the tokenization, classification, and tagging processes, consider the following requirement statement:

Req00: *The user must be able to display the PDF rendition of associated documents.*

The output of lexical analysis will be as follows:

Req00	User	must	be able to	display the PDF rendition of associated documents
Lbl	Ag	Mod	Const	Ac

E. Syntactic Analysis

The tokenized tagged requirement from the previous phase is input to the syntactic analysis phase. Syntactic analysis passes through the following process:

- 1) Reading each tokenized requirement.
- 2) Classifying the requirement into one of the six requirement classes depending on the modal phrase. The goal of this step is to decide which of the FSMs to use.
- 3) Using the suitable FSM to check whether the requirement statement follows the corresponding accepted syntax and to generate useful warning message as needed.

According to the final state the parser reaches, the user receives useful warnings as needed. Fig. 9 shows a screenshot of the tool depicting example input and output.

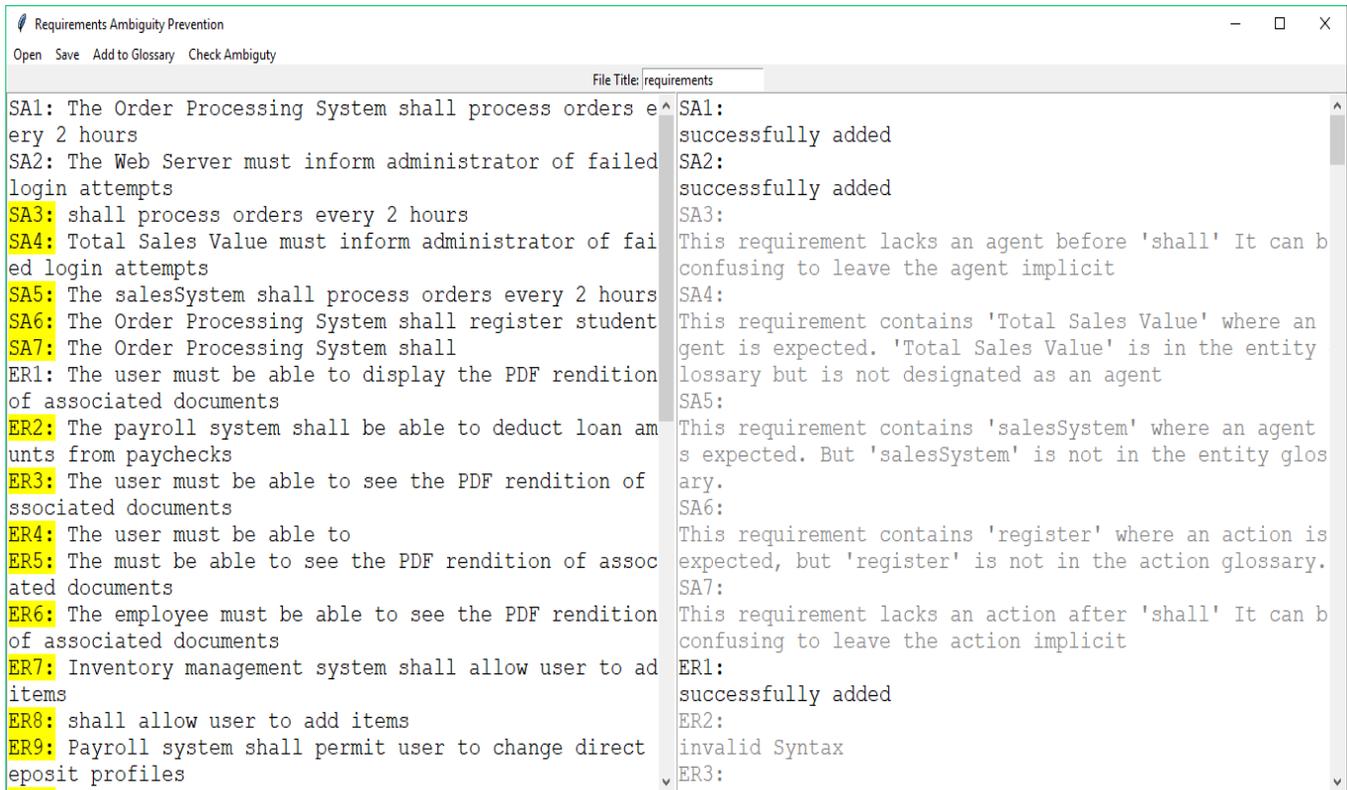


Fig. 9. A Screenshot of the Ambiguity Avoidance Tool; Example Input and Output.

IV. EXPERIMENT AND DISCUSSION

REFERENCES

A. Experimental Settings

We used Python version 3.6 to implement the ambiguity prevention tool. Then, we built a benchmark of 2460 real requirements. From the benchmark, we selected a random sample of forty real requirements. We classified each requirement in the sample into one of the six requirement classes mentioned above. We then transformed each classified requirement into the corresponding template and defined entities, agents, and actions in the glossaries. The purpose of this process is to emulate a real user writing the requirements before processing them through the tool.

B. Results and Discussion

From the experiment, it was clear that this approach can prevent some types of requirements ambiguity. Example issues that could be prevented using this approach are: missing information like missing an agent or missing an action; domain ambiguity like an unknown agent or an unknown entity; and non-best practices syntax like missing an action or an invalid syntax.

But on the other side, it was clear that classification and transformation processes are not straightforward. For example, some requirements had to be split into two requirements of different classes and templates.

Moreover, it was clear that the overall requirements writing process consumes more time and effort than using an uncontrolled natural language. In other words, there is a tradeoff between the effort needed to write the requirements following the predefined templates and ambiguity avoidance.

V. CONCLUSION AND FUTURE WORK

This paper presented details of a full implementation of a software requirements ambiguity prevention tool. This tool classifies the software requirements into one of six classes: solution, enablement, action constraint, attribute constraint, definition, or policy requirements. For each requirement class, there is an accepted defined template. To check whether the requirements adhere to the correct templates, the tool uses a FSM for each template.

We used Python to implement and test this approach. We selected forty random requirements sample out of 2460 real software requirements. We noted that the selected approach has some advantages and disadvantages as discussed above. But to judge this approach precisely, we need to compare it with other prominent approaches in our future work. It is important because we need to compare different approaches from some aspects such as: effectiveness in term of types and number of ambiguities resolved. We also need to compare the usability of the different approaches.

- [1] G. Sandhu, "Analysis of modeling techniques used for translating natural language specification into formal software requirements," *International Journal of Computer Applications*, vol. 113, no. 1, 2015.
- [2] H. Elazhary, "REAS: An interactive semi-automated system for software requirements elicitation assistance," *International Journal of Engineering Science and Technology*, vol. 2, no. 5, pp. 957-961, 2010.
- [3] K. Wiegers, "Karl Wiegers describes 10 requirements traps to avoid," *Software Testing & Quality Engineering*, vol. 2, no. 1, 2000.
- [4] K. Wiegers, "Writing quality requirements," *Software Development*, vol. 7, no. 5, pp. 44-48, 1999.
- [5] T. Stalhane and T. Wien, "The DODT tool applied to sub-sea software," in *2014 IEEE 22nd International Requirements Engineering Conference*, 2014, pp. 420-427.
- [6] C. Arora, M. Sabetzadeh, L. Briand, F. Zimmer, and R. Gnaga, "Automatic checking of conformance to requirement boilerplates via text chunking: An industrial case study," in *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2013, pp. 35-44.
- [7] B. Gleich, O. Creighton, and L. Kof, "Ambiguity detection: Towards a tool explaining ambiguity sources," *Requirements Engineering: Foundation for Software Quality*, pp. 218-232, 2010.
- [8] Y. Wang, I. L. M. Gutiérrez, K. Winbladh, and H. Fang, "Automatic detection of ambiguous terminology for software requirements," in *Natural Language Processing and Information Systems: Springer*, 2013, pp. 25-37.
- [9] K. D. Gill, A. Raza, A. M. Zaidi, and M. M. Kiani, "Semi-automation for ambiguity resolution in open source software requirements," in *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering*, 2014, pp. 1-6.
- [10] M. Bano, "Addressing the challenges of requirements ambiguity: A review of empirical literature," in *2015 IEEE 5th International Workshop on Empirical Requirements Engineering (EmpiRE)*, 2015, pp. 21-24.
- [11] P. Jain, K. Verma, A. Kass, and R. G. Vasquez, "Automated review of natural language requirements documents: Generating useful warnings with user-extensible glossaries driving a simple state machine," in *2nd India Software Engineering Conference*, 2009, pp. 37-46.
- [12] H. Elazhary, "Translation of Software Requirements," *International Journal of Scientific and Engineering Research*, vol. 2, no. 5, pp. 1-7, 2011.
- [13] C. Denger, D. M. Berry, and E. Kamsties, "Higher quality requirements specifications through natural language patterns," in *IEEE International Conference on Software: Science, Technology and Engineering*, 2003, pp. 80-90.
- [14] S. Farfeleder, T. Moser, A. Krall, T. Stalhane, I. Omoronyia, and H. Zojer, "Ontology-driven guidance for requirements elicitation," *The semantic web: Research and applications*, pp. 212-226, 2011.
- [15] Y. Wang, I. L. M. Gutiérrez, K. Winbladh, and H. Fang, "Automatic detection of ambiguous terminology for software requirements," in *International Conference on Application of Natural Language to Information Systems*, 2013, pp. 25-37.
- [16] K. T. Frantzi and S. Ananiadou, "Extracting nested collocations," in *16th Conference on Computational Linguistics*, vol. 1, 1996, pp. 41-46.
- [17] H. Yang, A. De Roeck, V. Gervasi, A. Willis, and B. Nuseibeh, "Analysing anaphoric ambiguity in natural language requirements," *Requirements engineering*, vol. 16, no. 3, p. 163, 2011.

A Comparative Study of the Decisional Needs Engineering Approaches

OUTFAROUIN Ahmad

Laboratory of Applied Mathematics
and Computer Science (LAMAI),
Faculty of Science and Technology
(FSTG), Cadi Ayyad University,
Marrakech, Morocco

ZAHID Nouredine

High Teachers School
Cadi Ayyad University
Marrakech, Morocco

ABDALI Abdelmounaïm

Laboratory of Applied Mathematics
and Computer Science (LAMAI),
Faculty of Science and Technology
(FSTG), Cadi Ayyad University,
Marrakech, Morocco

Abstract—Requirements Engineering (RE) is an important phase in a project of systems development. It helps design-analysts to design and to model the expression of the end-user needs, and their expectations vis-a-vis their future system. This engineering is studying two major issues that are: What should the system do in order to have a complete needs specification, and reason on the why: "Why do we need to build this system? ", without looking for how to build it. The vast majority of needs engineering approaches are based on two concepts: scenario or goal; there are generally three types of approaches: Scenario-Oriented Approaches, Goal-Oriented Approaches and approaches generated by the couple: goals and scenarios at the same time. In the remainder of this paper, we present a comparative study of the three types of the RE approaches, then models of needs representation, and finally we conclude with the conclusions.

Keywords—Decisional information systems; decisional needs engineering; needs engineering approaches; goal; scenario; model of needs representation

I. INTRODUCTION

Today, decisional information systems have become indispensable to help in making the decision. According to earlier studies, about 60% of the errors in the projects of system development come up during the Requirements Engineering (RE) phase [1], a relatively young field of research: until the end of the 80s was still referred to as "analysis" to qualify the upstream phase of system design; the analysis phase is essential to produce the specifications of the system to be developed.

Needs Engineering (NE) was introduced by J. Hagelstein [2] and E. Dubois [3] to designate the part of the development of information systems that concerns the investigation of users' problems and needs, and the development of the future system specifications. It helps to express what the system has to do, but not how it should do it. Moreover, in order to have a complete specification of needs, we must, also, reason on the why: "Why do we have to build the future system? ".

In classical information systems, the RE was presented by Rolland and al. [4] as a process that derives from requirements through the exploration of the objectives of the actors and the activities to achieve them, and in the decisional domain, we talk about the Decisional Needs Engineering (DNE) which is

defined according to Nuseibeh and al. [5] as a discipline that takes care of : elicitation, analysis, specification, validation and management of needs and constraints for the construction of a system phases.

Several approaches have been proposed to analyze the decision-makers' needs, this approaches are oriented by a process formed by a set of phases which are broken down into a set of steps (Fig. 1), this process is accompanied by models of representation of the needs during the analysis phase.

In the formalization of decision-making needs (DN), the vast majority of DNE approaches are based on the following concepts: goal or scenario. These two concepts are the source of three types of approaches: Scenario-Oriented Approaches, Goal-Oriented Approaches and approaches generated by the couple: goals and scenarios at the same time.

The remainder of this paper is a DNE approaches comparative study, using the following structure: Section 2 presents a comparative study of the approaches to the DN analysis. In Section 3, we discuss a state of the art of the decisional needs representation models. This work will be completed by conclusions in Section 4.

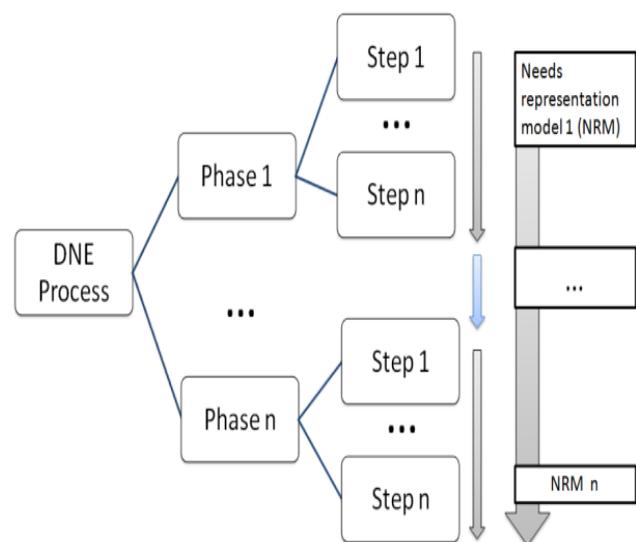


Fig. 1. Decomposition of DNE Process.

II. NEEDS ENGINEERING APPROACHES

The success of a NE project relies heavily on the success of the NE process, which typically consists of the following phases:

- Elicitation of the needs: A phase that helps to understand the organizational situation and the expression of needs [6], [7], [8], [9], [10].
- Specification: Defines the relation between the business objective and the functional and non-functional components of the system [11], [12], [4], [13].
- Negotiation: The phase in which we define the deliberation context of the whole process [14], [15].
- Validation: phase of validation of the system specifications with regard to the needs expressed / expected by the users [16], [17].

To ensure the quality of this process, it is essential to have appropriate techniques, approaches and tools; the choice of these three elements influences the quality of the resulting needs.

In the next section, we first look at the Goal-Oriented Approaches. Next, we cite the Scenario-Oriented Approaches. Finally, we review the approaches that combine goals and scenarios.

A. The Goal-Oriented Approaches

According to Ben Achour [18], a goal is defined as "Something that someone hopes to achieve in the future". We find in other works that the goal can be defined as "an objective to be achieved in the future system" [19]. In other words, a goal is an image of an intention, which is subsequently operated on by a set of objectives that are planned to be realized in a precise duration, without specifying how they can be reached, it is associated with a result that we want to have and materialize by a set of object states.

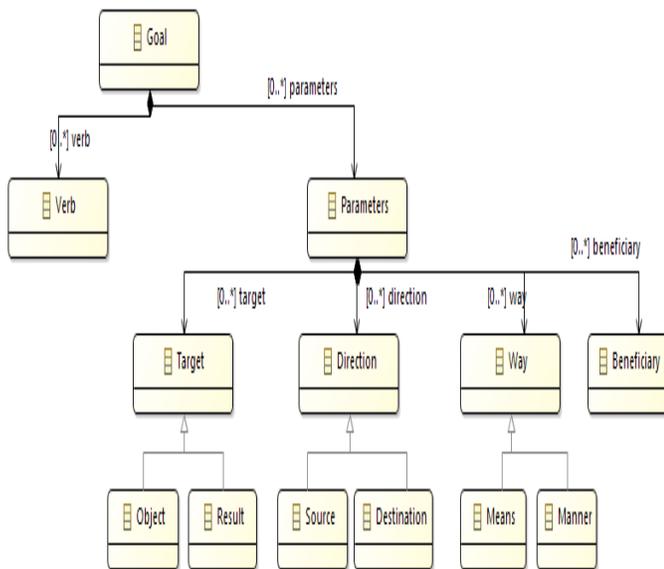


Fig. 2. Structure of a Goal [Prat, 1999].

1) *Structure of a goal*: In general, the goal is expressed in natural language, and formalized according to a structure composed of a verb accompanied by a set of parameters, each of them has a semantic function and provides in their instances answers to the different questions that are Around this verb: who, what, when, how much, how etc.

This structure is proposed at the beginning in the works of Prat [20] (Fig. 2) which in turn relies on the grammar of the cases of Fillmore [21] and on its extensions. This goal structure is subsequently improved in other works [22], [23].

In this structure, we have mandatory components to define: the verb and the target, but the other parameters are optional:

- **Target**: The target is a complement to the action concerning the entities affected by the goal. There are two types of targets: the object and the result. The object exists before achieving the goal and may, eventually, be modified or deleted by the goal; whereas the result is the entity resulting from the realization of the goal designated by the action.
- **Quantity**: it measures the quantity of the object that should be produced.
- **Quality**: This is a property that must be achieved or preserved.
- **Direction**: Contains two types of directions named: source and destination, their role is to identify, respectively, the initial and final locations of the object:
 - ✓ **Source**: Represents the starting point of the goal (source of information or physical location).
 - ✓ **Destination**: Represents the ending point of the goal (to whom or to what).
- **Beneficiary**: Expresses the person or group for whom the goal should be obtained.
- **Way**: It consists of two parameters:
 - ✓ **The manner**: Specifies how the goal can be achieved.
 - ✓ **The means**: Specifies by what means (tool) can the goal be achieved.
- **Locality**: It positions the goal with regard to space.
- **Time**: It positions the goal with respect to time.

Reference: it is the entity according to which an action, of the fact table, is performed or a state is achieved or maintained.

The advantage of using natural language is to simplify and to facilitate the manipulation of these intentions, which are represented in the form of a linguistic formulation, and their understanding by the different actors / participants in the Decisional Information Systems (DIS) and more particularly in the process of the RE. In Elgoli's work [22], she was inspired by this linguistic formulation and she proposed a new version in the form of a meta-model (Fig. 3) expressing the semantics and facilitating automatic exploitation while remaining understandable by the actors.

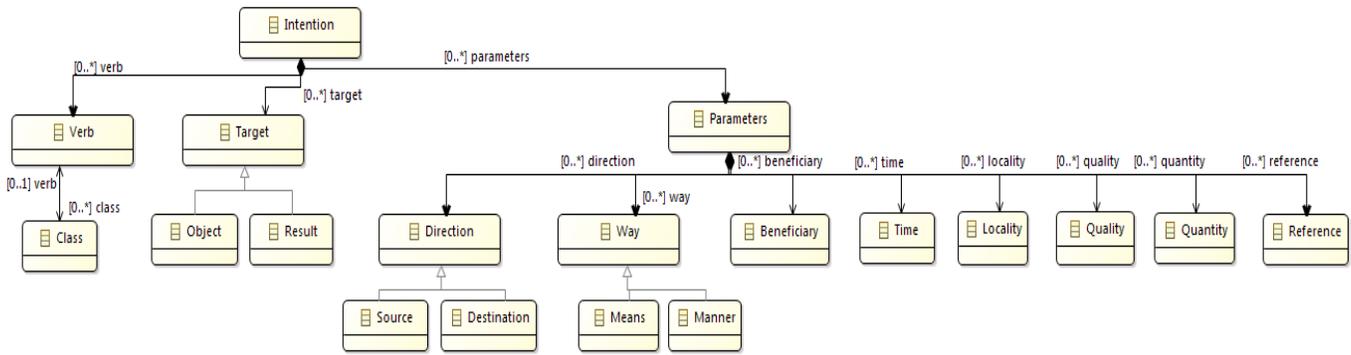


Fig. 3. Linguistic Meta-Model of Intention in UML Notation [ELGOLLI, 2008].

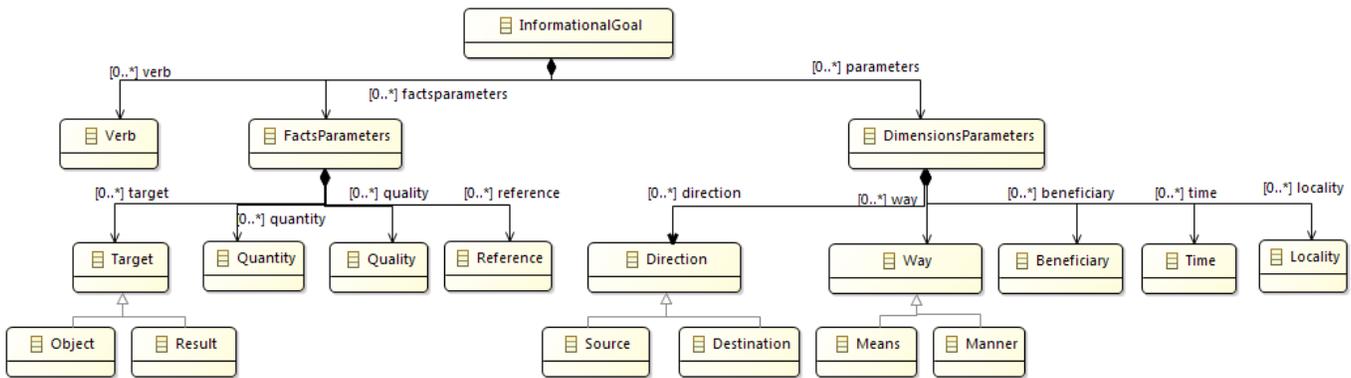


Fig. 4. Semantic Model Proposed to Represent an Informational Goal [23].

By following the same approach, Sabri [23] in her work extended this work of Elgoli [22] by trying to make appear the facts' parameters and the dimensions' parameters at the moment of the semantic representation of the informational goal (Fig. 4). To facilitate the way for the operational actors of the DIS to develop the decision data dictionary on which we will base to build the multidimensional star schema.

2) *Levels of goal abstraction*: In the decisional field, a strategic goal (level 1) does not offer an operational view and must be decomposed into tactical goals, this level (Level 2) does not yet give us the possibility to deduct our facts and our dimensions; thus we move on to the third level (level 3), which is operational, by dividing each tactical goal into a set of informational goals (Fig. 5).

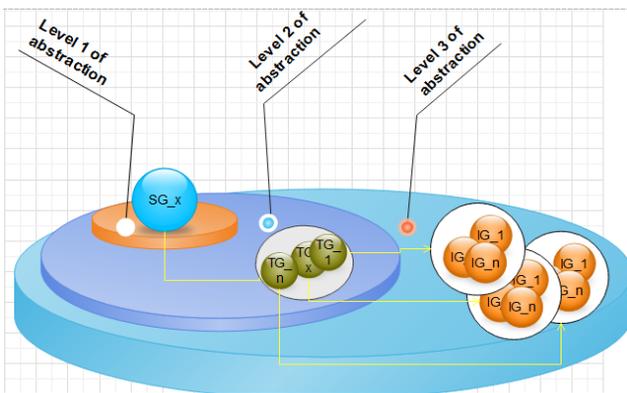


Fig. 5. Levels of the Goal Abstraction.

Therefore, each decisional need (n) is decomposed into a set of strategic goals (SG) and each strategic goal i is presented as a set of tactical goals (1 to n), thus:

$$DN = \sum_{i=1}^n SG_i$$

Such as :

$$SG_i = \sum_{j=1}^n TG_j$$

And for every tactical goal j of the strategic goal i, it is, itself, presented as a collection of informational goals (from 1 to m), we have:

$$TG_j = \sum_{k=1}^m IG_k$$

In DNE's approach, DNs are classified according to these levels of abstraction. Hence the classification of decision-making goals into three categories: strategic, tactical and informational.

The treatment and decomposition of a goal into sub-goals has been studied in several works that we decompose according to three categories:

The first category: We use AND / OR [11], [24], [25], [26] and [27] reduction graphs which have inspired this method of artificial intelligence [28].

A goal 'A' can be decomposed into several sub-goals: A1, ... An.

If an AND relation is associated with goals {A1 AND A2}, {A1 AND A3} ... implies that all of these goals must be achieved to achieve the desired result of goal A, and that one cannot replace the other.

In this case the satisfaction of one goal (A1 for example) ensures the satisfaction of the other (A2).

The second category: several approaches have extended the method used in the first category, with some variations from one approach to another; we find those that have adopted a new hierarchical organization of goals based on the relations AND, OR and REFINED BY [29], this last link "Refined by" is deduced when the two goals share a syntactic part of the goal and are complementary, but do not aim at the same result.

In other works [23], another type of "complemented" link is used to represent a particular case of the OR relation; this link is used to express a relation between two goals that share a syntactic part and the two syntaxes are complementary and aim to achieve the same result.

The third category: It is a contribution that we have proposed in our work [30]. To facilitate its treatment, each goal is decomposed into a result and a canal; the result is decomposed into a set of actions and the canal is decomposed into a set of means and a set of manners.

$$\text{We have : Result} = \sum_{i=1}^n \text{Action } i$$

Canal = $\sum_{i=1}^n \text{Means } i + \sum_{j=1}^m \text{Manner } j$, we present this relation in a class diagram (Fig. 6).

The analyst-designer can represent the links between the goals of the same type by the relations' matrices between the {Strategic / Tactical / Informational} goals. The link is a combination of {R: same Result, -R: Different result} And {C: same Canal, -C: Different canal}, after this matrix it is treated with a set of rules that have already been developed according to the possible cases [30].

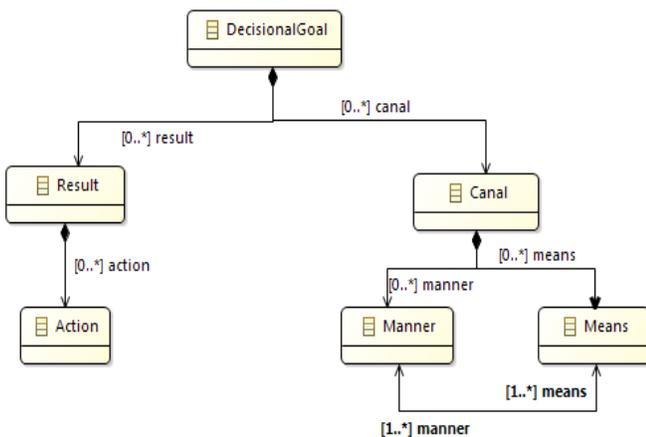


Fig. 6. Meta-Model of a Decisional Goal.

B. The Scenario-Oriented Approaches

Scenarios have been used to capture user needs [31] [32]. According to Rolland [18] a scenario is defined as "a possible behavior limited to a set of interactions between several agents". It allows achieving a given goal by interacting two agents. It is also a way of describing the different behaviors and the different perspectives that the actors wish to have or expect in relation to the use of their system. Hence, each scenario is characterized by its state, its result and a set of conditions likely to influence the behavior of the agents.

According to Rolland [33], a scenario can be presented as the "order of actions or events for a specific case of a certain generic task that a system must perform". Otherwise, it represents, in a comprehensible way, a sequence of events and activities (which are collected according to the needs of the various actors involved in the design of the system) Connected in a conditioned manner in order to achieve a result or realize a functionality.

1) *Language structure of a scenario*: In principle, natural language is used to represent the set of actions that constitute the scenario; these actions are chained, between an initial state and a final state, according to conditions.

Several structures are proposed to model the scenarios; in the work of Tawbi [34], he proposed a new linguistic model of a scenario (Fig. 7), based on that defined by Ben Achour [18].

In this model, Tawbi considers two states for the scenario: Initial and Final and distinguishes two types of scenarios: normal scenarios and exceptional scenarios. A normal scenario achieves the desired result of the goal, while an exceptional scenario ends with the non-satisfaction of the goal. The actions are of two types: atomic and flux. An atomic action is an interaction between two agents affecting an object. An agent and an object can appear in several different interactions. A flow of actions is used to define the scheduling between interactions in a scenario. It is composed of several actions. Action's flows are classified into four types: sequence, competition, repetition or constraint.

In Rolland's work [33], we define the scenario with four axes: its form, content, purpose and its cycle (Fig. 8).

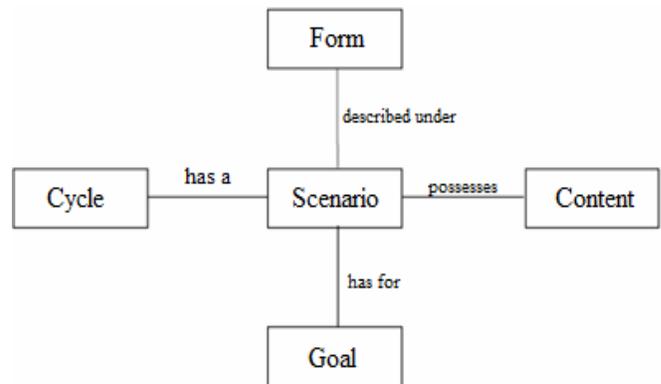


Fig. 7. Scenarios' Aspects [Rolland, 1998].

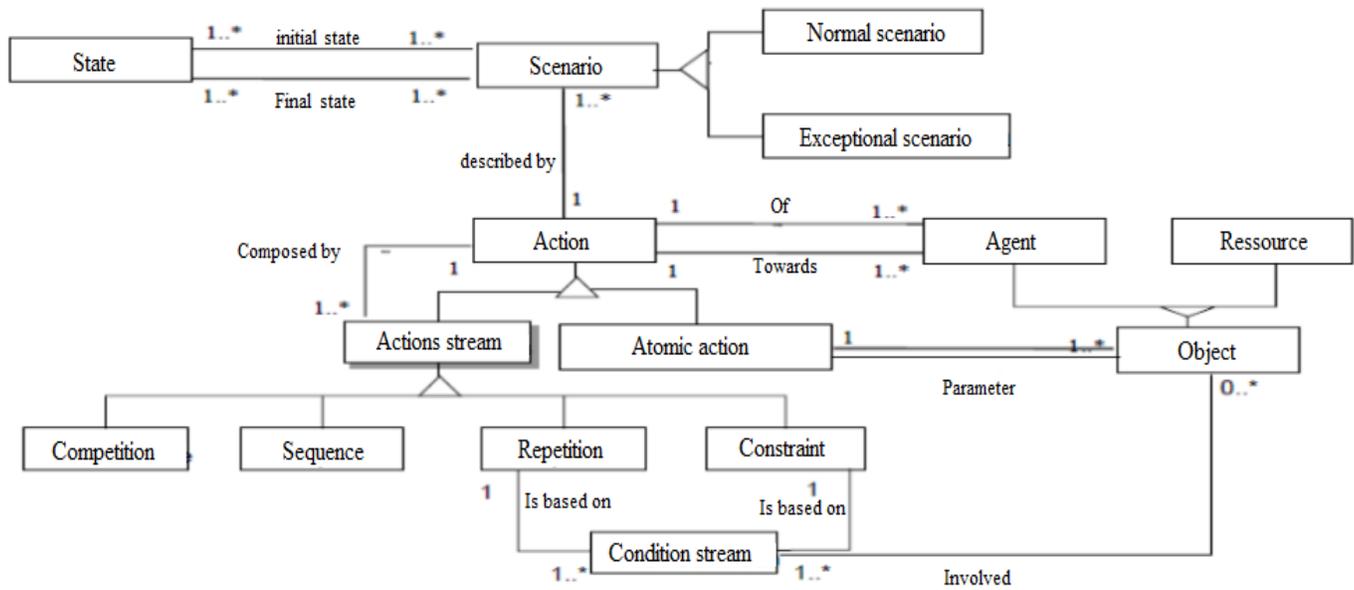


Fig. 8. Structure of a Scenario [Tawbi, 2001].

a) *Form*: The form of a scenario is very important during the acquisition, specification and representation of needs and goals in order to validate or evaluate them.

In the table (Table I) we have established a study of some description forms of the scenarios used in several methods.

For the description of a scenario, mainly three notations are found: informal (using natural language that is sometimes more appropriate for users), semi-formal [35] (based on structured notations such as tables [32] or The scripts [36]) or formal (scenarios are represented with languages based on regular grammars [37] or state diagrams [38], UML forms (scenarios are represented by sequence diagrams or collaboration diagrams), as well as other forms the automaton [39, 40], statecharts [41, 37], Formalisms derived from the Petri nets [42, 43, 44], etc.).

TABLE I. SCENARIOS' FORME

Methods \ Scenarios' forme	Jac92	Hsi94	Kos94	GH95	Kaw97	Lus97	Dan97	Rol98	Elk98	Lee98	Kh-99	Elk00
Sequence Diagrams or Collaboration Diagrams (UML)	×											×
Regular grammars		×										
Automaton					×	×						
Statecharts			×	×							×	
Formalisms derived from Petri nets							×		×	×		

This description of the scenarios is made, on the one hand, to simulate the different functionalities that the future system must have, and on the other hand, to link the reactions of the users who will trigger them.

b) *Content*: The content refers to the type of information and knowledge that the scenarios will contain (Table II).

The content also depends on how the scenarios describe the system. There are abstract scenarios that refer to abstract objects: Customer, Provider. And concrete scenarios that refer to concrete objects that are particular instances of the object: Faculty of Science and Technologies (FST), University Cady Ayyad (UCA).

TABLE II. SCENARIO CONTENT CHARACTERISTICS

Method	Content characteristics
[Jacobson, 1996]	<ul style="list-style-type: none"> Describe the internal functioning of the system Describe the interaction between the system and its environment Describe the organizational aspects of the system
[Dardenne et al., 1993]	<ul style="list-style-type: none"> Describe the functional aspect (structure, behavior, system functions) Describe the non-functional aspect (organizational consideration, performance, risk management) Describe the intentional aspect (goal-oriented approaches, accountability-based approaches)
[Kyng, 1995]	<ul style="list-style-type: none"> Describe organizational levels Describe strategic levels [45]

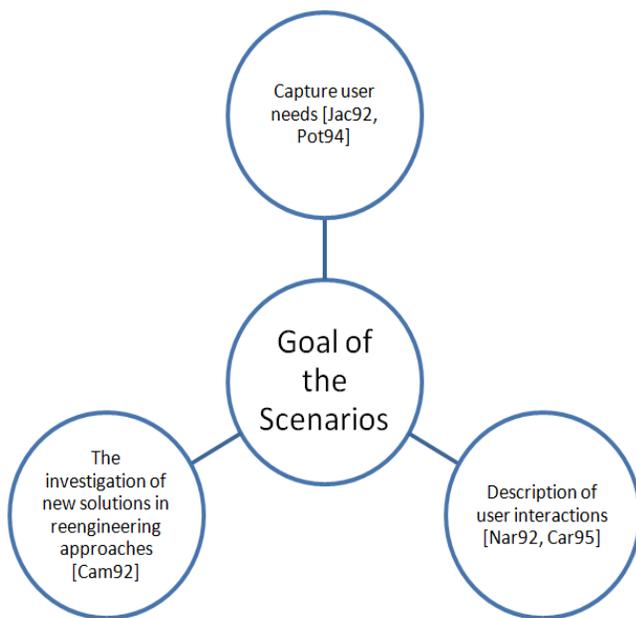


Fig. 9. Scenarios' Goal.

c) *Goal*: The goal of scenarios is usually one of three things (Fig. 9):

- **Description**: These are scenarios that describe the behavioral aspects of the system by representing the views of external users to the system [32], [46].
- **Exploration**: this type of scenarios is used to choose the best solution among many that is explored and evaluated [47], [48].
- **Explanation**: This is a type of a scenario that is used after exploratory scenarios in order to defend and explain the details of the chosen solution [49].

d) *Scenario life cycle*: Considering the scenarios being objects describing the system, we have two types:

- **Persistent scenarios**: According to Jacobson [31] and Potts [32], scenarios accompany the project from the needs' analysis to the production of the documentation.
- **Temporary scenarios**: In contrast to the first type, these scenarios intervene just in certain stages of the development cycle of a project (e.g. scenarios for the acquisition of needs, or for their validation [50]).

In the scenarios, we distinguish between a normal scenario and an exceptional scenario (a scenario that describes what happens if the normal scenario does not work) [8], but the limit always remains in describing what will not happen in exceptional cases and is not a real-life scenario since it does not help in achieving the goal. In our work [51], we proposed a new formalization of the associating for each goal two scenarios: normal(NS) and alternate(AS) which form both a set of steps. Each step in the PS can have its alternation in the AS, so that we are on to have a mechanism to reach our goal

before we begin our decision-making project by trying to avoid all problems of this kind.

2) *Goal-Scenario directed approaches*: The objective of these approaches is to discover the needs of the system by coupling each discovered goal with a scenario that illustrates the behavior of the system to achieve the goal.

A goal is "intentional" while a scenario is "operational". Therefore, is possible to combine the two concepts. Each goal can be attached to one and only one scenario (which operationalizes it), and each scenario describes the steps and constraints of achievement (describes a possible behavior of the system to achieve the goal) of one and only one goal. The couple <goal, scenario> is named a fragment of need [33] and explains a part of the specification of the system to be realized. The fragments of needs can be classified at various levels of abstraction: the contextual level to which the services rendered by the system in the context of the organization are identified, the level of interaction in which the behavior of the system is described and the interactions which must carry out with its users and the physical level in which the behaviors of the internal objects of the system are described.

This approach was evaluated through four different experiments:

a) Workshops [33]

b) Case study [29]: Four characteristics that contribute to the satisfaction of the discovery of the needs of the system: 1) The notion of fragment of need is defined as the couple <goal, scenario>. 2) The hierarchical organization of needs is based on the relations AND, OR and refined by, between fragments of needs. 3) The elicitation process is based on a bidirectional movement between a goal and a scenario. For a given goal, a scenario is written to illustrate its realization. 4) A methodological help, in the form of semi-automatic rules, is implemented by the software The Ecritoire.

c) Empirical studies [52],

d) CREWS-Ecritoire project [34].

The results obtained by these experiments validated the applicability and effectiveness of this approach. The ECRITOIRE approach [53] is the software application of the CREWS approach [34]. It interprets and transforms a scenario to ensure its consistency, completeness and conformity to the goal. It proposes: 1) methodological guidelines for writing textual scenarios (written in natural language) and software tools to check their correction; 2) scenario analysis rules helping to discover variants; Exceptions and complements of a given scenario, and (3) a formalization of the process while guiding its development.

III. MODELS OF THE NEEDS' REPRESENTATION

Each step of the DNE approaches corresponds to the establishment of a model, which facilitates the capitalization and archiving of the DNE process. The models of the representation of the requirements are classified into five categories of models (Table III):

TABLE III. COMPARATIVE STUDY OF THE MODELS OF THE NEEDS' REPRESENTATION.

Methods Needs' representation	Lujan- Mora and al., 2003 [56]	Ghozzi, Ravat and al., 2005 [60]	Mazon, Trujillo and al., 2005 [55]	Feki, Ben Abdallah and al., 2006 [59]	Annoni, 2007 [54]	Gam El Golli, 2008 [22]	Bargui, Feki and al., 2009 [58]	Abdelhédi and Zurfluh, 2013 [57]	Sabri Aziza and al, 2015 [23]
Goal models			×		×	×			×
Query models		×						×	
Table models							×	×	
Models based on relational schemas	×			×					

- Goal models: Numerous studies are based on the "i *" goals' model [3], which is a modeling language; it is defined with the dependencies between various types of agents, in order to model situations where one of the agents depends on another to achieve a certain goal, or to carry out a task. Other works [4] propose a method for analyzing the decision-makers' needs using a goal model to represent the intentions and the implemented strategies to achieve a goal.
- Table models: The collection of the decision-makers' needs in the table models is made via n-dimensional tables containing the concepts of : facts, dimensions, measurements, parameters, hierarchies and attributes. To collect the needs, we ask decision-makers to express them in a syntactic model [5]. Afterwards, the analyst-designer extracts and treats the multidimensional concepts and generates multidimensional schemas.
- Models based on relational schemas: The formalization of decisional needs is made by several types of relational schemas, such as the Entity / Association model [6]. The authors use an ideal schema for the formalization, from which we define a candidate schema for the treatment phase; it is on the basis of this schema that our conceptual schema is generated.
- Query models: Queries, in this kind of approaches, are the basis of the modeling of decisional needs. Initially, the expressed requirements are captured in natural language from which the analyst-designer formalizes these needs in the form of queries. The next phase of needs' treatment, in which we extract fact indicators (fact table and its measurements) and dimension indicators (dimension tables and their attributes) is done with a matrix of needs [7]. After this step, we define the first star schema extracted using the needs and we confront it with a second star schema which will be made using the data sources.
- Mixed models: In this category, two or more types of models are combined in order to collect, formalize and treat needs. For example, needs can be collected in the form of queries and subsequently be formulated into goals and into decisions. The authors use an owner goals' model GDI (Goal / Decision / Information) to represent them [8]. In other works [9] to treat DNs, a

model of analytical requirements' specification is used (queries / tables) to extract fact tables and dimension tables.

IV. CONCLUSION

In this paper, we have made a comparative study of engineering needs approaches and classified them into three categories: goal-directed approaches, scenario-based approaches, and approaches mixed goals and scenarios.

We also studied the structure of a goal and a scenario, the formalization of a study and the study of models of needs' representation. These concepts are the starting point for defining and organizing the needs of designers of models, which allows us to establish an intentional level of abstraction to facilitate the reuse of the modeling process and tools.

Goal-directed approaches, generally provide goal modeling by decomposition into the form of trees and/or. Scenario-driven approaches derive conceptual models from scenarios and are used to reason about design choices. Finally, in mixed approaches, the scenarios are used to describe different possible ways of achieving the same goal, so the goals are operationalized by the scenarios.

In the future work we will define a new modelization of decisional need, based on the goal levels of abstraction, we will define a new more relevant axes of goals treatment with new treatment rules and a new formalization of the informational goals to facilitate the extraction of indicators on fact tables (with its measurements) and indicators on their dimension tables (with their attributes) associated.

ACKNOWLEDGMENT

I acknowledge the support provided by my supervisors: Pr. Abdelmounaim ABDALI and Pr. Noureddine ZAHID and the members of the laboratory LAMAI (Laboratory of Mathematics Applied and Informatics) of the Faculty of Science and Technology-Cadi Ayyad University Marrakesh.

REFERENCES

- [1] B. Boehm, "Software Engineering Economics", Prentice Hall. 1981.
- [2] J. Hagelstein, "Declarative approach to information systems requirements", Knowledge-Based Systems, vol.1, n°4, pp.211-220, 1988.
- [3] E. Dubois, J. Hagelstein, and A. Rifaut, "Formal Requirements Engineering with ERAE ", Philips Journal Research, vol.43, n°4, 1989.

- [4] C. Rolland, and N. Prakash, "Bridging the Gap Between Organisational Needs and ERP Functionality", *Requirements Engineering Journal*, vol.5, n°3, pp.180-193, 2000.
- [5] B. Nuseibeh, and S. Easterbrook, "Requirements Engineering : A Roadmap", In *Proceedings of International Conference on Software Engineering*, ACM Press, Limerick, Ireland, 4-11 June 2000.
- [6] S. Card, T. Moran, and A. Newell, "The Psychology of Human-Computer Interaction", Lawrence Erlbaum Associates, Hillsdale, NJ, USA, ISBN: 0898592437, 1983.
- [7] C. Ellis, and J. Weaine, "A conceptual model of groupware", CSCW'94, Chapel Hill, NC, 1994.
- [8] V. Kavakli, and P. Loucopoulos, "Goal-driven business process analysis application", in *electricity deregulation Information Systems*, 24(3):187-207, 1999.
- [9] E. Yu, "Agent Orientation as a Modelling Paradigm", *Wirtschafts informatik*, 43(2), pp.123-132, 2001.
- [10] H. Bendjenna, "Ingénierie des Exigences pour les Processus Interorganisationnels", PhD thesis, University Mentouri of Constantine (Lab. LIRE) and the university of Toulouse (EDMITT, Lab. IRIT), 21 November 2010.
- [11] A. Dardenne, A. Van Lamsweerde, and S. Fickas, "Goal directed requirements acquisition", *Science of Computer Programming*, 20 (1-2), pp.3-50, 1993.
- [12] A. I. Anton, "Goal based requirements analysis", *Proceedings of the 2nd International Conference on Requirements Engineering ICRE'96*, pp.136-144, 1996.
- [13] L. Chung, B. Nixon, E. Yu, and J. Mylopoulos, "Non-Functional Requirements in Software Engineering", Kluwer Academic Publishers, 2000.
- [14] B. Ramesh, and D. Vasant, "Supporting systems development by capturing deliberations during requirements engineering", *Software Engineering*, IEEE Transactions on 18.6 (1992):498-510, 1992.
- [15] J. Lee, "Design Rationale Systems: Understanding the Issues", *IEEE Expert Intelligent Systems and Their Applications*, 12 (3):78-85, 1997.
- [16] V. R. Basili, "Applying the Goal/Question/Metric paradigm in the experience factorySoftware", *Quality Assurance and Measurement: A Worldwide Perspective*, ISBN-10: 1850321744, Edition "Intl Thomson Computer Pr (Sd) (June 1995)", pp.21-44, 1993.
- [17] S.P. Wilson, T.P. Kelly, and J.A. McDermid, "Safety Case Development: Current Practice, Future Prospects", in *Proceedings of 1st ENCRESS/5th CSR Workshop*, September 1995.
- [18] C. Ben Achour, "Extraction des Besoins par Analyse des Scénarios Textuels", PhD thesis, University Paris 6, Paris, France, Jan. 1999.
- [19] V. Plihon, J. Ralyté, A. Benjamen, N.A.M. Maiden, A. Sutcliffe, E. Dubois, and P. Heymans, "A reuse-oriented approach for the construction of scenario based methods", *Proceedings of the International Software Process Association's 5th International Conference on Software Process (ICSP'98)*, Chicago, Illinois, USA, June 14-17, 1998.
- [20] N. Prat, "Réutilisation de la trace par apprentissage dans un environnement pour l'ingénierie des processus", PhD thesis, University Paris1, France, 1999.
- [21] C.J. Fillmore, "Lexical Entries for Verbs", *Foundations of Language*, Vol. 4, No. 4 pp. 373-393, Nov., 1968.
- [22] I. Gam El Golli, "Ingénierie des Exigences pour les Systèmes d'Information Décisionnels : Concepts, Modèles et Processus (la méthode CADWE)", PhD thesis, University Paris-Panthéon-Sorbonne, France, October 2008.
- [23] A. Sabri, and L. Kjiri, "Une approche d'Ingénierie des Besoins Décisionnels pour la conception d'Entrepôts de Données dans un contexte de réutilisation", PhD thesis, University Mohammed V of Rabat (ENSIAS, Rabat), Morocco, March 28, 2015.
- [24] X. Bubenko, C. Rolland, P. Loucopoulos, and V. De Antonellis, "Facilitating 'fuzzy to formal' requirements modeling", *IEEE 1st Conference on Requirement Engineering, ICRE'94*, pp.154-158, 1994.
- [25] P. Loucopoulos, and V. Karakostas, "Systems Requirements Engineering", McGraw-Hill, London, UK, 1995.
- [26] J. Mylopoulos, K.L. Chung, and E. Yu, "From object-oriented to goaloriented requirements analysis", *Communications of the ACM*, vol.42, n°1, pp.31-37, 1999.
- [27] A. Van Lamsweerde, "Requirements Engineering in Year 00: a Research Perspective", *Proc 22nd international Conference on Software Engineering*, Limerick, 2000.
- [28] N. J. Nilsson, "Problem Solving Methods in Artificial Intelligence", McGraw Hill, 1971.
- [29] C. Rolland, G. Grosz, and R. Kla, "Experience With Goal-Scenario Coupling In Requirements Engineering", *Fourth IEEE International Symposium on Requirements Engineering (RE'99)*, University of Limerick, Ireland, 7-11 June 1999.
- [30] A. Outfarouin and A.Abdali, "On a new modeling process of the decision-makers' needs", *IJCSNS International Journal of Computer Science and Network Security*, Vol.17, No.2, February 2017.
- [31] I. Jacobson, M. Christenson, P. Jonsson, and G. Oevergaard: "Object Oriented Software Engineering: a Use Case Driven Approach", Addison-Wesley, 1992.
- [32] C. Potts, K. Takahashi, and A.I. Antón, "Inquiry-based requirements analysis", *IEEE software* 11, n°2 (1994):21-32, 1994.
- [33] C. Rolland, C. Souvey, and C. Ben Achour, "Guiding goal modelling using scenarios", *IEEE Transactions of Software Engineering*, Special Issue on Scenario Management, vol.24, n°12, December 1998.
- [34] M. Tawbi, "CREWS-L'Ecritoire : un Guidage Outillé du Processus d'Ingénierie des Besoins", PhD thesis, University of Paris 1-Sorbonne, France, 2001.
- [35] K. Weidenhaupt, K. Pohl, M. Jarke, and P. Haumer, "Scenario usage in system development : a report on current practice", *IEEE Software*, March 1998.
- [36] K. S. Rubin, and A. Golberg, "Object Behavior Analysis ", *Communications of the ACM*, 35(9), pp.48-62, September 1992.
- [37] M. Glinz, "An Integrated Formel Model of Scenarios based on Statecharts", In *Fifth European Software Engineering Conference*, Lecture Notes in Computer Science, Vol.989, pp.254-271, Springer-verlag (1995).
- [38] D. Harel, "Statecharts : a Visual Formalism for Complex Systems ", *Science Computer Program* 8, pp.231-274, 1987.
- [39] I. Kawashita, "Spécification Formelle de Systèmes d'Information Interactifs Par La Technique de Scénarios", Master thesis, Université de Montréal (1997).
- [40] F. Lustman, "A Formal Approach to Scenario Integration", *Annals of Software Engineering*, Vol. 3, pp.255-272 (09/1997).
- [41] K. Koskimies and E. Makinen, "Automatic Synthesis of State Machines from Trace Diagrams", *Software Practice & Experience*, Vol.24, No.7, pp.643-658 (1994).
- [42] B. Dano, H. Briand and F. Barbier, "An Approach based on the Concept of Use Case to Produce Dynamic Object-Oriented Specifications", In *proceeding of the Third IEEE International Symposium on Requirements Engineering*, pp.54-64, Annapolis, 1997.
- [43] M. Elkoutbi and R. K. Keller, "Modeling Interactive Systems with Hierarchical Colored PETRI Nets", In *Proc. of 1998 Adv. Simulation Technologies Conf.*, pp.432437, Boston, MA (04/1998).
- [44] W.J. Lee and Y.R. Kwon, "Integration and Analysis of Use Cases Using Modular Petri Nets in Requirements Engineering". *IEEE Transactions on Software Engineering*, Vol.25, No.12, pp.1115-1130 (12/1998).
- [45] M. Kynig, "Creating Contexts for Design", In *Caroll J.M. editor, Scenario-Based Design : Envisioning Work and Technology in System Development*, pp.85-107. John Wiley and Sons (1995).
- [46] R. Guillerme, N. Sadou, and H. Demmo, "ESA Petri net: Dynamic reliability analysis Tool", *International Journal of Adaptive and Innovative Systems*, vol.1, n°3/4, pp.201-216, 2010.
- [47] C. H. Holbrook, "A scenario-based methodology for conducting requirements elicitation", *ACM SIGSOFT, Software Engineering Notes*, vol.15, n°1, pp.95-104, 1990.

- [48] K. Allenby and T. Kelly, "Deriving Safety Requirements using Scenarios", 5th IEEE International Symposium on Requirements Engineering (RE'01), IEEE Computer Society Press, 2001.
- [49] P. Wright, "What's in a Scenario", ACM SIGCHI Bulletin, vol.24, n°4, October 1992.
- [50] P. Hsia, J. Samuel, J. Gao, D. Kung, Y. Toyoshima, and C. Chen, "Formal Approach to Scenario Analysis", IEEE Software, Vol.11, No.2, pp. 33-41, Mar. 1994.
- [51] A. Outfarouin and al., "Towards a new decisional needs formalization." Computer Systems and Applications (AICCSA), 2016 IEEE/ACS 13th International Conference of. IEEE, 2016.
- [52] M. Tawbi, F. Velez, C. BenAchour, and C. Souvey, "Scenario Based RE with CREWS-L'Ecritoire: Experimenting the approach", ESQ'2000, Sixth International Workshop on Requirements Engineering: Foundation for Software Quality, Stockholm, Sweden, June 5-6 2000.
- [53] C. Rolland, "L'Ingénierie des besoins : l'approche l'ECRITOIRE", published in Journal Techniques de l'Ingénieur, Paris, France, 2003.
- [54] E. Annoni, "Eléments méthodologiques pour le développement des systèmes décisionnels dans un contexte de réutilisation", PhD thesis, University of Toulouse 1, Toulouse, France, 2007.
- [55] J.-N. Mazon, J. Trujillo, M. Serrano, and M. Piattini, "Designing data warehouses: from business requirement analysis to multidimensional modeling", Proceeding of the 13th IEEE International Requirements Engineering Conference Workshop on Requirements Engineering for Business Needs and IT Alignment (REBNITA), Paris: August 2005.
- [56] S. Luján-Mora, and J. Trujillo, "A comprehensive method for data warehouse design", Proceeding of the 5th International Workshop on Design and Management of Data Warehouses, DMDW'03, Berlin, Germany, September 2003.
- [57] F. Abdelhédi, and G. Zurfluh, "User Support System for Designing Decisional Database", ACHI 2013: The Sixth International Conference on Advances in Computer-Human Interactions, Nice, France: 24 Feb. - 1 Mar. 2013.
- [58] F. Bargui, J. Feki, , and H. Ben-Abdallah, "A natural language approach for Data Mart schema", NLDB'09: Proceedings of the 14th international conference on Applications of Natural Language to Information System, Saarland University, Saarbrücken, Germany: 23-26 June 2009.
- [59] J. Feki, H. Ben-Abdallah, and M. Ben-Abdallah, "Réutilisation des patrons en étoile", INFORMATIQUE des ORGANISATIONS et SYSTÈMES d'INFORMATION et de DÉCISION (INFORSID 06), 31 mai- 4 june, Hammamet, Tunisie, 2006.
- [60] F. Ghazzi, F. Ravat, , O. Teste, , G. Zurfluh, "Méthode de conception d'une base multidimensionnelle contrainte", Revue des Nouvelles Technologies de l'Information – Entrepôts de Données et l'Analyse en ligne (EDA'05), Cépadués éditions, volume RNTI-B-1, pages 51–70, 2005.

A Blockchain Technology Evolution between Business Process Management (BPM) and Internet-of-Things (IoT)

Doaa Mohey El-Din M. Hussein, Mohamed Hamed N. Taha, Nour Eldeen M. Khalifa
Faculty of Computers and Information
Cairo University
Egypt

Abstract—A Blockchain is considered the main mechanism for Bitcoin concurrency. A Blockchain is known by a public ledger and public transactions stored in a chain. The properties of blockchain demonstrate in decentralization as distribution blocks, stability, anonymity, and auditing. Blockchain can enhance the results of network efficiency and improve the security of network. It also can be applied in several fields like financial and banking services, healthcare systems, and public services. However, the research is still opening at this point. It includes a big number of technical challenges which prevents the wide application of blockchain, for example, scalability problem, privacy leakage, etc. This paper shows a proposed comprehensive study of blockchain technology. It also examines the research efforts in blockchain. It presents a proposed blockchain lifecycle which refers to an evolution and a linked ring between business process management improvement and Internet-of-Things concepts. Then, this paper presents a practical proof of this relationship for smart city. It presents a new algorithm and a proposed blockchain framework for 38 blocks (which recognized as smart-houses). Finally, the future directions are well presented in blockchain field.

Keywords—Blockchain; bitcoin; business process; cryptography; decentralization; consensus; applications

I. INTRODUCTION

A business process refers to a collection of related tasks to achieve the delivery process about service or product. Business process management (BPM) is keen on the design, execution, monitoring, and improvement of business processes. Systems include main four processes: analysis, design, enactment, and execution of the used processes in companies to streamline and automate intra-organizational processes. BPM is interested in improving corporate performance through managing the business processes [1].

The blockchain is a type of business process management but it makes a revolution in the management of processes as a study in optimization execution [2]. Blockchain technology can be a solution to interoperability, trust, and transparency issues in divider networks or systems. At its core, blockchain is a distributed ledger of asset and transaction records.

The term blockchain includes two threads a network and a data structure. Blockchain has a big difference from distributed database in data integrity. No one can update or delete any record in blockchain business process that will be a

benefit and secure in many fields as healthcare or hospitals profiles about patients.

Blockchain is considered one of data structure unit/system [3], a blockchain includes a linked list of blocks, each containing a set of transactions. The crypto-graphical is the main property of each block in the chain. The data structure is duplicated across a network of instruments. Each instrument carrying the full replica is known a full node.

Blockchain is also a network when it is authorized a combination of peer-to-peer networks, consensus-making, cryptography, and market techniques. Blockchain's name that came from the data structure (fact) which a chained list of blocks. This chain is spread as a peer-to-peer network, in which every node retains the final updated version of it. This is considered that the immutable of blockchain history is very usefully.

The blockchain does not require to any authority reverse a central mechanism. It relies on a distributed node for sharing the data on the network. A consensus is required to achieve it on the network for each participant. In Bitcoin concurrency application, there is a function to deduce a consensus that is called a Proof of Work function [4]. This strategy requires that any node wishing to add a block to the blockchain must complete a computationally expensive (but easily verifiable) puzzle first. In healthcare applications, a patient can give the proof for the access of his information.

The Internet of things (IoT) [5] is a physical network for any resources can be controlled remotely as machines, devices, home appliances, and other items embedded systems. Each provenance had to make a software, sensors, and connectivity which enables these things to connect and change data. The development of intelligent objects is considered a distributed and decentralized ledger technologies with smart contracts. The internet of things (IoT) has an effective role in management and disruption these business practices. The main concepts must be taken into this process are security and accountability. They impose several rules with government regulations as finical systems. That causes a reduced costs and risk, reduce time delays, improved quality and consistency.

By integrating IoT [6] with decentralized blockchain technology that combines smart contracts, BPM and BPO will

enter a world where just about anything can be connected to communicate intelligently.

This paper shows the relationship between blockchain and business process and internet of things. This evolution of relationship deduces main challenges are faced when constructing blockchain. The rest of this paper is organized as follows: Section 2 shows the related works. Section 3 is a presentation of Blockchain-Based on BPM & IoT. In Section 4, outlines of the Blockchain advantages. Section 5 highlights the Blockchain challenges. In Section 6, the outlines of a case study on the smart city. In Section 7, the outlines of discussion. Finally, Section 8 the conclusion and future research direction.

II. RELATED WORK

This section presents a summary of the essential aspects of blockchain technology [7] and discusses initial research efforts at the intersection of BPM and IoT.

A. Business Process Management (BPM)

BPM is a business philosophy about people, and how to work together, and the performance objectives in their process. Business process management has a sequenced workflow including automate, monitor and analysis [2]. There are several distributions of lifecycle as in the following that has analysis, design, execution, implementation, monitoring and adaption.

The BPM lifecycle [8] differs as the organization or application but no one can exclude the main process analysis, design, execution and implementation. This lifecycle enables to apply the management system.

This paper illustrates several works on BPM and lifecycle, conditions, rules and structures.

The authors [9] presented a system for healthcare workflow in two hospital environments. They created the analytical framework based on six theoretical propositions identified as having a major impact on the implementation of workflow technologies.

In [10], researchers examined BPMS and presented a help in conditional structures such as if, switch, and while. They made a comparison between implementation of common conditional structures. This comparison relied on their workflow management systems via case studies, as well as discusses capabilities of each system.

In [11], the authors leaded to a BPMS research study. They discussed some details about workflow-related concepts and their typologies, references of some BPMS and current research trends and hotspots.

B. Blockchain-Technology

The blockchain [12, 13, and 14] is an original invention that can't be denied known by the pseudonym, Satoshi Nakamoto. A blockchain technology generated the backbone of a new type of internet. Bitcoin is an original digital concurrency which is based on blockchain technology. Blockchain [15] relies on four attributes and concepts: distributed shared ledger, cryptography, consensus, smart contracts.

Blockchain came from the fact of its chained list of blocks as data structure. This chain of blocks distributes over a peer-to-peer network, in which every node maintains the latest version of it. Blocks can contain information about transactions. Blockchain, the technology implicit Bitcoin, is a kind of Distributed Ledger Technology that has been known as a “distributed, shared, encrypted database that serves as an irreversible and incorruptible repository of information” [16].

A block includes a header and the body. The header of each block contains:

- Block version: refers to the rules of validation based on a set of blocks.
- Parent block hash: contains a 256-bit hash rate.
- Merkle tree root hash: the hash value of all the transactions in the block.
- Timestamp: refers to the timestamp by seconds currently.
- nBits: includes the existing hashing target in a compact format.
- Nonce: a 4-byte field, which usually starts with 0 and increases for every hash calculation.

Fig. 1 illustrates the blockchain structure and its components [17].

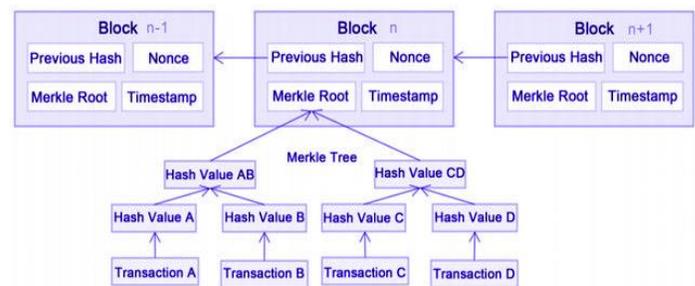


Fig. 1. The Blockchain Structure

The transaction counter and transactions are the two parts of the block body. The ultimate number of transactions for each block relies on the size of the block and the size of each transaction. A cryptography mechanism is used asymmetrically in blockchain to support and confirm the transactions' authentication [11]. The Digital cryptographic signature is utilized in an untrustworthy environment [12].

In the moment of the entered information, it could not be erased or updated. Blockchain is both the network and database in security and the data integration [13].

According to, the blockchain is a software communicator that can support making explicitly important architectural considerations on the resulting performance and quality attributes of the system. The authors' research [14] supported the architectural decision on whether to appoint a decentralized blockchain as opposed to other software solutions, such as traditional shared data storage. Additionally, they examined specific implications of the usage of blockchain as a software

connector containing design trade-offs regarding quality attributes.

A semantic layer built upon a basic blockchain infrastructure would join the benefits of flexible resource/service discovery and validation by consensus. The researchers presented a novel Service-oriented Architecture (SOA) semantically [15].

The authors present BLOCKBENCH which is considered a benchmarking framework for performance perception of private blockchains versus data processing workloads. A study made a comprehensive evaluation of three major blockchain systems according to BLOCKBENCH [21] is entitled Ethereum, Parity and Hyperledger Fabric. The results are illustrated several trade-offs in the design space, as high-performance gaps between blockchain and database systems. Drawing from design principles of database systems, they were examined several research directions for bringing blockchain performance closer to the realm of databases [16].

C. Business Processes Improvement by using Blockchain Technology

Blockchain is like a service that enables a business to leverage all the advantages of cloud computing while the blockchain assessment and implementation: flexibility, agility, capex-free, compliance, scalability, pay as you go, allowing business to deal with decentralized blockchain network concurrently. A time can be saved in creating blocks, managing a blockchain network, designing the network, and the developing applications, swiftly pool and validate use cases. Rapidly scale and roll out blockchain based services.

1) Blockchain Levels

There are three levels of Blockchain [17]: Storage for digital records, Exchanging digital assets, and Implementing the smart contracts requires knowing the basic rules, and understanding terms, properties and conditions recorded for the implemented code. Distributed network performs contract & monitors compliance. The results are evaluated without the third party automatically.

2) Blockchain Types

The types of blockchain include three formative mechanisms as in Fig. 2:

a) *Public: Everyone can check the transaction and verify it.*

b) *Consortium: It refers to the node that had authority can be chosen in advance, usually has partnerships like business to business, the data in blockchain can be open or private, can be seen as Partly Decentralized.as Hyperledger.*

c) *Private: it refers to not every node can participate, maybe one or more restricted in the constructed blockchain. That interpretation of the firm authority for the access of data management.*



Fig. 2. Types of Blockchain

3) Blockchain & BPM Applications

In addition, a private-public key mechanism coupled with powerful cryptographic algorithms keep everything secure. Blockchain applications damage conventional thinking and conventional ways according to the processing of data, handling, and storage.

The inter-organizational processes used blockchain represent [2]: the control flow as a big part in blockchain and business logic of its processes can be executed from the process models into the blockchain smart contracts. That is known trigger components allowed connecting these inter-organizational process implementations to Web services and internal process implementations. These triggers can build a bridge between the enterprise applications and the technology of blockchain. The cryptocurrency basic can enable the selective implementation of conditional payment and built-in escrow management at defined points within the process, where this is required and feasible to clarify these capabilities. This may very well be a basis for misunderstandings and shifting blame in cases of conflict [2, 3].

The technical realization of this advance is still nascent at this stage, although some early efforts can be found in the literature. For example, smart contracts that implement the trust execution process from BPMN process models [2] and from domain-dependent [18]. Further, the evaluation of optimizations costs is presented by [3].

The previous examples presented that blockchain technology and how to apply on BPM application. It is important to pass the real technical issues blend with promising application scenarios; early implementations mix with unanticipated challenges.

D. Blockchain in IoT and Blockchain Applications

Internet of things (IoT) [5] is considered a worldwide network of interconnected objects and human beings, which through singular addressing schemes are able to interact with each other and participate with their neighbors to reach common targets [19].

The primary purpose of IoT is to share objects and entities information that examines the manufacture, transportation, and other specifics of people's life. Through the Iot information, it could produce a preferable cognitive and environment. But the development of the IoT is still slowly these years. IoT has a lifecycle for applying in any domain, as in the following:

This lifecycle is faced several problems in blockchain technology. That may be affected on the time and confidence. Because when the blockchain applies that required approved from all blocks to any updates.

One important reason is that the high costs of the deployment. The security and privacy of Internet of Things (IoT) exist major challenge because of the big scale and distributed nature of IoT networks. The approaches of blockchain technology serve the decentralized security and support the privacy. So far, they comprise significant energy, delay, and overhead that is not appropriate for most resource-constrained IoT devices.

The essential logistics basics are traceability and transparency. IBM Blockchain makes optimization business transactions and trading relationships with substantially secure business networks on blockchain—both at scale and globally.

Blockchain companies attracted \$ 525 million in 2015, largely as a result of a peak in investments in the first quarter of 2015. Investments then decreased until the fourth quarter of 2015 when only \$ 45 million was invested 67% of the fundraising activities carried out between January 1, 2015 and February 18, 2016 concern companies specializing in Blockchain infrastructure and applications. The remaining 33% are companies specialized in bitcoin [20]. Fig. 3 illustrates the final sectors percentage using blockchain technology.

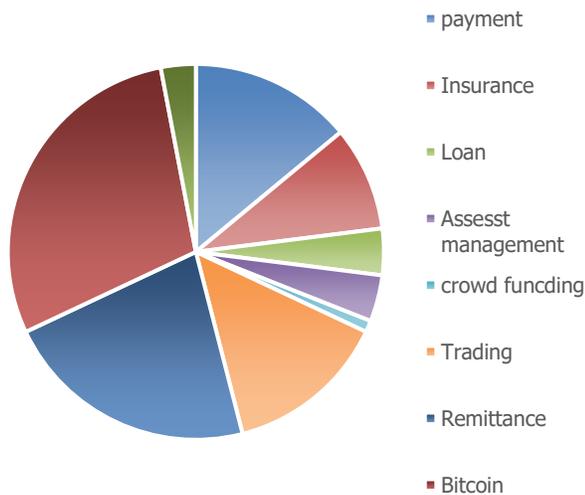


Fig. 3. Financial Sectors used Blockchain Distributions

This section discusses applications and examples for using blockchain in IoT.

1) Electric Power Microgrids:

Electric power can be used blockchain in different block sizes as big or small. Smart contracts are being used for redistributing excess power from solar panels. The Transactive Grid is an application running on blockchain to monitor and redistribute energy in a neighborhood microgrid. The program presented saving costs and reducing pollution through buy and

sell processes automatically. The technology for running the program is the Ethereum platform, designed for building smart contracts of any kind [21].

2) Cold Chain Monitoring

Food and pharmaceutical products mostly want a specific pilling. Also, enterprises also see the value in sharing warehouses and distribution centers, instead of each one paying for its own. Sensors on sensitive products can record temperature, humidity, vibration, and other items of interest.

These readings can then be stored on blockchain. They are permanent and tamperproof. If a storage condition deviates from what has been agreed, each member of the blockchain will see it. A smart contract can trigger an action to correct the situation. Depending on the size of the deviation, this action may be to simply adjust the storage. However, it could also extend to changing “use-by” dates, declaring products unfit, or applying penalties [22].

3) Meat Traceability

Product status at each stage of production can be recorded using blockchain. The records are permanent and inalterable. They also allow the tracing of each product to its source. Global retailer Walmart uses blockchain to track sales of pork meat in China. Its system lets the company see where each piece of meat comes from, its processing and storage, and sell-by date. In the event of product recall, the company can also see which batches are concerned and who bought them [23].

4) Automotive Supplier Payments

Blockchain allows the transfer of funds anywhere in the world. Bitcoin transfers specifically also offer lower fees. Australian vehicle manufacturer Tomcar uses bitcoin to pay suppliers. The advantage is in the cost savings. On the other hand, the firm is careful to avoid hanging onto too much bitcoin. While bitcoin is international by nature, some national governments see it as a way for companies to make an investment. Companies with bitcoin holdings may, therefore, be taxed accordingly [24].

They presented [25] a lightweight instantiation of a BC particularly geared for use in IoT by eliminating the Proof of Work (POW) and the concept of coins. The authors’ research was exemplified in a smart home setting and consists of three main tiers namely: cloud storage, overlay, and smart home. They presented that Blockchain smart home system is secure by thoroughly analyzing its security. They introduced results of their simulation and the overheads.

III. COMPERHENSIVE STUDY

This paper presents a comprehensive study between ten research papers about blockchain related to business process management (BPM) and Internet-of –things (IoT) as in Table I.

IV. BLOCKCHAIN-BASED IN BPM AND IOT

There are also challenges and opportunities for BPM and blockchain technology beyond the classical BPM lifecycle. We refer to the BPM strengths [26] beyond the methodological support we reflected above, including strategy, governance, information technology, people, and culture.

TABLE I. THE COMPREHENSIVE STUDY BETWEEN TEN RESEARCHES RELATED TO BLOCKCHAIN AND BUSINESS PROCESS

Paper No.	BP challenge	Technique used	Domain	Goal
[4]	Risk of centralized data is the security of taking footprint and requires centralized trust in a single authority	Block chain in three algorithms: Creating blocks Proof of interoperability Miner election.	Health	Describe an approach to effectively and securely share healthcare information within a data sharing network
[26]	Risk Adaptations	The companies Study works on blockchain technology based	Financial	Study improvement of blockchain in business as Bitcoin concurrency efficiency as Visa, Mastercard, Banks, NASDAQ, etc., are investing in exploring application of current business models on Blockchain.
[27]	-T capabilities and Infrastructure -Information and cyber security -Integration and collaboration	Supply chain	Delphi study's expert	portraying the emerging transition trend from a digital business environment, the presented Delphi study findings contribute to extant knowledge by identifying 43 opportunities and challenges linked to the emergence of Big Data Analytics from a corporate and supply chain perspective
[13]	Scalability problems: storage optimization of blockchain re-designing blockchain	Blockchain testing could be separated into two phases: standardization phase and testing phase	Organization	They presented a Comprehensive survey on blockchain including blockchain architecture and key characteristics of blockchain.
[21]	Performance evaluation , bridging Database Design into data model layer for BC, and Scalability	BlockBench	YCSB small bank	BlockBench measures overall and component-wise performance regarding throughput, latency, scalability, and fault-tolerance
[16]	the adoption of blockchain in the supply chain and logistics	Financial and non-financial spheres	Industries	Consider the possible adoption of blockchain-based application, created by the Finnish company Kouvola Innovation.
[20]	Performance evaluation for semantic-enhanced blockchain	Logistics , industry, Utility markets , Public sector, Financial services	several domains of smart cities and communities	Improve performance for Semantic based blockchain enhancement in different domain
[23]	the unambiguous and correct specification of smart contracts	the structure of Nested ADICO (nADICO)	domain-specific language	It can automate the translation of institutional constructs into codified machine-readable contractual rules.
[15]	Scalability and performance of The main bottlenecks in Hyperledger and Ethereum are the consensus protocols	BLOCKBENCH	a domain name registrar	Improving blockchain performance
[6]	blockchain-IoT combination	distributed peer-to-peer systems	Several industries	It makes blockchain-IoT combination to facilitate the sharing of services and resources

And the strengths [28], if we use blockchain in IoT, are low cost, Flexible system, higher Security, Systematic, and high efficiency. We find the new role to can combine between the two concepts of BPM and IoT, how to enter IoT in blockchain lifecycle according to manage a business process.

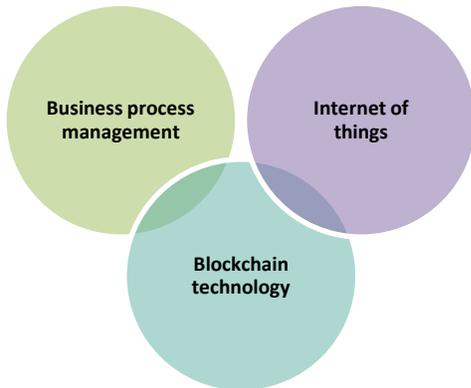


Fig. 4. The relationship between blockchain to BPM & IoT.

This relationship will reflect on blockchain lifecycle and differs in the processes sequences or importance. Blockchain technology raises another relevant perspective for estimating high-level processes in terms of the implied strengths, weaknesses, opportunities, and threats.

The proposed lifecycle merges between the main characteristics and how to affect each process in the next process in business process management. As a result of the Fig. 4, IoT can serve in management systems and be a system high secure, that shows the relationship between blockchain to BPM and IoT. This lifecycle has several challenges as scalability, security, complexity, confidentiality, and domain dependent.

A. The Blockchain Lifecycle

The blockchain lifecycle includes 7 processes. These processes are:

1) *The Analysis*: is concerned to acquisition insights into issues relating to the way a business process currently operates. Each organization can apply the blockchain processes analysis shared internally and externally by stakeholders [28]

2) *IoT security people*: Blockchain Enforces rules BPM-related information technology subsumes all systems that support process execution, such as process-aware information systems and business process management systems. These systems typically assume central control over the process. Blockchain technology might modify governance to be more externally depending on smart contracts [29].

3) *Design* [30, 31]: which refers to the identification and distribution of the blocks on the network. It also determines network type (private, public or hybrid models). And the security and authority for each block through the network.

Fig. 5 illustrates the proposed lifecycle of relationship in blockchain and BPM & IoT 4.

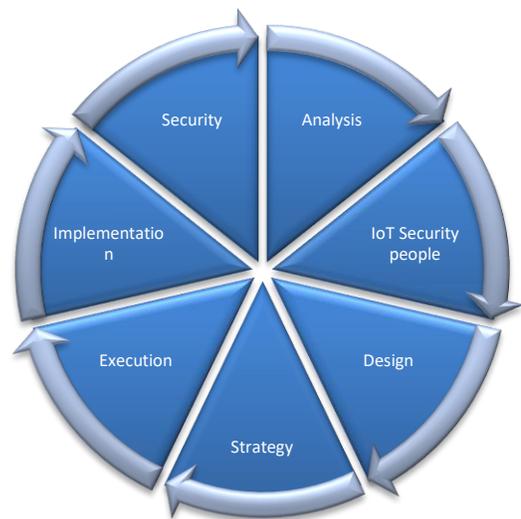


Fig. 5. The proposed lifecycle of relationship in blockchain and BPM & IoT.

4) *Execution*: that is keen on the instantiation [32, 33]of individual cases and their information-technological processing. Recently, it is important to aware each process in the information systems and in the management systems [34]. The essential challenge of the implantation engineering level is the identification and definition of abstractions for the design of blockchain and processes execution.

5) *Implementation*: refers to the procedure of transforming a to-be model into software components executing the business process [34]. In this context, the question is how can the involved parties make sure that the implementation that they deploy on the blockchain supports their process as desired. Some of the challenges regarding the transformation of a process model to blockchain artifacts are discussed by [35].

6) *Monitoring Process*: is supportive events for implementing and executing way, and triggering alerts to identify strange behavior [36].

B. Organizational knowledge

Organizational knowledge is defined by the collective values of a group of people in an organization [13]. Currently, BPM is discussed in relation to organizational culture [37] from a perspective that emphasizes an affinity for clan and hierarchy culture.

Blockchain privacy and Governance refers to appropriate and transparent accountability in terms of roles, responsibilities, and decision processes for different BPM-related programs, projects, and operations [3]. Currently, BPM can define the roles of BPM and properties in each organization internally. Blockchain technology provides a governance model orientation. Research on corporate governance investigates agency problems and mechanisms to provide effective incentives for intended behavior. Smart contracts can be used to establish new governance models as exemplified by The Decentralized Autonomous Organization (The DAO).

V. BLOCKCHAIN TECHNOLOGY ADVANTGES

The blockchain can improve the security and transparency through all kinds of transactions, put the probabilistic of the possibilities. It also can represent at the supply chain [16]. This technology enhances in the tasks:

- Recording the quantity and transfer of assets - like pallets, trailers, containers, etc. – as the nodes movement of supply chain.
- Tracking purchase orders, change orders, receipts, shipment notifications, or other trade-related documents
- Assigning or verifying certifications or certain properties of physical products; for example determining if a food product is an organic or fair trade.
- Linking physical goods with digitalizing numbers or codes.
- Sharing information.

Blockchain offers shippers the following advantages:

- Improved Transparency. Documented products
- Greater Scalability. Virtually any number of participants, accessing from any number of touchpoints, is possible.
- Better Security. It can share the data in ledger to audit the requirements and support company internally.
- Increased Innovation. Opportunities abound to create new, specialized uses for the technology as a result of the decentralized architecture.
- Confidant-Ability
- Data integrity

VI. BLOCKCHAIN TECHNOLOGY CHALLNEGES

There are several challenges [38, 39] in blockchain but this paper focuses on the challenges when blockchain mechanism intervenes in BPM and IoT concepts. The proposed lifecycle faces challenges when anyone tries to build it. These challenges are shown as scalability, security, complexity, speed cost and domain dependent.

1) *Scalability Blockchains*: This refers to the big size of blockchain network and grows continuously. That may cause of several attacks on this network. So that requires often distributed management systems and powerful network to harvest any problem on the network.

2) *Trade-off Transaction Costs and Network Speed*: According to several political aspects when using bitcoin, there is a problem of store information and how to get miners the information or rerecord them.

3) *The blockchain Complexity*: It has made cryptography more mainstream. The recent researches support several types of glossaries and searching indexes to make it easy to understand.

4) *Confidentiality*: Mute information in a circle and do not change it to protect the data.

5) *Based on specific-Domain*: Each blockchain relies on one structure of domain with basic rules and conditions of it. No one till now can enter several data about different topics to cover different topics in each block. That is a problem in management data and resources.

6) *Security*: The still open research problem insecurity illustrates in the lying or rumor review. If more than a number of half working persons on the network say one lying that will be a true fact.

For this reason, the mining of bitcoin pools are demonstrated carefully by the community, to guarantee no strange people on the network. That means the Politics in blockchain: the protocols of the blockchain present a chance to digitize governance models. Another reason, miners are ultimately forming another kind of incentivized governance model, there have been ample chances for public collisions between various community strips.

There are other challenges to deal with the blockchain technically:

1) *Mindset*: Blockchain was targeted for the decentralized research. There is still a problem in mind thinking in the idea of the centralization for the most students or researchers in blockchain orientation.

2) *Human Error*: If a blockchain is used as a database, the information going into the database needs to be of high quality. The blockchain stores data unstructured, so that requires to evidence when registering the data correctly in first computer/block.

3) *Technology And Know-How*: Blockchain programming takes a mix of software skills. It also helps to understand economies and businesses, especially your business. You may have to train staff or hire new people with these skills. The development of blockchain outsourcing that can support the third party.

VII. EXPERIMENT STUDY

If executed using smart contracts on a proposed life cycle of blockchain, typical barriers complicating the deployment of smart-city.

- The blockchain supports a fixed public ledger so the entrants can trust the history writing messages to define the error source.
- Smart contracts can monitor the independent process globally. That interparty only predicated messages are agreeable.
- The data will be encrypted for visible in public blockchain network. These capabilities examine the useful blockchains for communities and organizations to perform the business and organizations' boundaries.

This is an essential improvement, because the blockchain core uses to provide enterprise collaborations going far beyond asset management, raise safety or sharing the personal smart

city records in purchase (home) block providers as in the following, which we propose this algorithm.

In this experiment, we design small simulation architecture in smart cities. That targets higher security system for each owner/ node in the city and easier to sell and buy process from the trusted owners. Fig. 6 shows the proposed experiment of Blockchain in smart city.

Blockchain also can control automated systems for several owners/ users in each building. Blockchain technology relies on distributed decentralized ledgers. The simulation works on thirty-eight computers and virtual computers (distributed devices and databases) in a lab to test the blockchain concept.

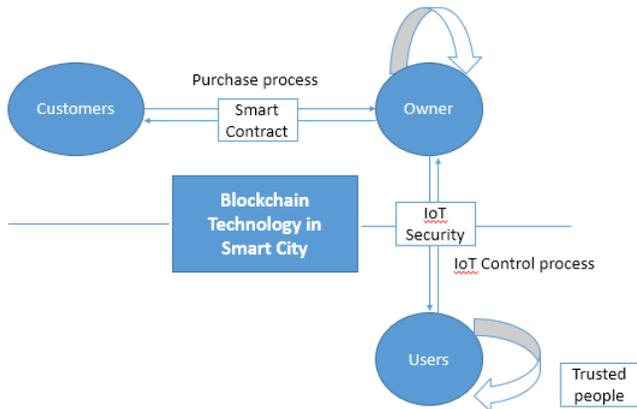


Fig. 6. Blockchain in Smart city

1) Contract Signing Stage

Contracts are signed by electronic data interchange (EDI) in traditional E-business [40], and digital signature is used to guarantee the legal effect. Other insurance measures include negotiation logs and files. This electronic evidence can be used to arbitrate the legal dispute in the transaction. But these evidences have to be kept in a server that is managed by a third party.

2) Contract Fulfillment Stage

This stage starts after both the buyer and seller have completed all the procedure of the contract. In the classical E-business, the seller should be ready for customers' requirements in goods and issues by the evidence, insurance and available credit then the delivery process for goods with transporting an organization to finish rest tasks as packing, shipping and transportation. The buyer should transmit the money of goods through banks accounts after confirming buying process of these goods. Even if these smart contracts are preserved by some companies or organizations, there is no guarantee that the content of them will not be modified or delete by someone. Therefore, we need to publish the smart contract into the Blockchain. Here are two transaction styles in IoT E-business. One is the payment, the other is the exchange. The formal one can be applied to the purchase of the commodities and services on the IoT.

Table II presents the proposed Blockchain Algorithm in Smart City. In other words, this table can discuss the smart housing for Building City Internally. For example, one people

want to buy commodities from a DAC. For example, B want to buy a car from A. First, Smart contract includes the exact terms of the transaction. Second, both sides confirm this contract and publish it into the Blockchain. Lastly, the contract will take effect and both sides will get what they need.

TABLE II. THE PROPOSED BLOCKCHAIN ALGORITHM IN SMART CITY (SMART HOUSING FOR BUILDING CITY INTERNALLY)

```

1. Function PurchaseBuilding (id)
2.   if msg.value == (Vid)   Vid → valide building
3.   if customer == Tid     Tid → trusted customer
4.     customer= Trust
5.     bdata =GetBuilding data (BuildingUrl)
6.     if bdata exist and bdata is valid then
7.       send value to customer
8.       return bdata
9.     if (customerAccount > =bdata) && (customerCase="oK")   ok →
      accept to purchase
10.      send bitcoins to ownerAccount
11.      return "purchase successfully"
12.    else (customerAccount >=! bdata) && (customerCase="oK")
13.      return "Account not allowed"
14.    else
15.      return "Not interested to buy"
16.  else
17.    Send value to owner
18.    return "Building data is invalid"
19.  end if
20. end if
21. end if
22. end function
    
```

VIII. CONCLUSION AND FUTURE WORK

This paper presents a comprehensive study of blockchain technology and its effect in business process management and Internet-of-thing. It presents IoT life cycle and its relation between BPM lifecycle. It also proposes a solution for higher security in blockchain in a smart-city. Blockchain can transform supply chains, industries and ecosystems. Smart contracts can save time and cost and eliminate any delay. So that make the chain is faster and more intelligent and higher security of supply chain.

REFERENCES

- [1] Book chapter: Introduction to Business Processes, BPM, and BPM Systems, Springer International Publishing Switzerland 2015 A. Burattin: Process Mining Techniques in Business Environments, LNBP 207,
- [2] Mendling, J., Weber, I., Van Der Aalst, W., Brocke, J. V., Cabanillas, C., Daniel, F., ... Zhu, L., Blockchains for Business Process Management - Challenges and Opportunities. ACM Transactions on Management Information Systems, 9(1), 2018.
- [3] Luciano, G., Alexander, P., Marlon, D., and Ingo, W., Optimized execution of business processes on blockchain, BPM'17: International Conference on Business Process Management , Barcelona, Spain, 2017.
- [4] Kevin, P., Rammohan, D., Pradip, K., and Kelly, B., Mayo C., A Blockchain-Based Approach to Health Information Exchange Networks, 2016
- [5] Dave, E., The Internet of Things How the Next Evolution of the Internet Is Changing Everything, Cisco Internet Business Solutions Group (IBSG), 2011.
- [6] Konstantinos, C., & Michael, D., Blockchains and Smart Contracts for the Internet of Things, 2016 IEEE. Translations and content mining are permitted for academic research only, volume 4, 2016.
- [7] Allan, T., and John, D., Linked Data Indexing of Distributed Ledgers, WWW '17 Companion Proceedings of the 26th International Conference on World Wide Web Companion, Australia 2017.

- [8] Stefan R. Koster, An evaluation method for Business Process Management products, University of Twente, Netherlands, 2009.
- [9] R.S. Mans¹, M.H. Schonenberg¹, M. Song¹, W.M.P. van der Aalst¹, and P.J.M. Bakker², Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital, A. Fred, J. Filipe, and H. Gamboa (Eds.): BIOSTEC 2008, CCIS 25, pp. 425–438, 2008., springer 2008.
- [10] Emir M. BahsiEmrah CeyhanTevfik Kosar, Conditional Workflow Management: A Survey and Analysis, *Dblp, Scientific Programming* 15(4):283-297, 2007
- [11] A. Meidan n , J.A. García-García, M.J. Escalona, I. Ramos, A survey on business processes management suites, *ACM*, 2017
- [12] Haoyan Wu ¹ , Zhijie Li ¹ , Brian King ¹ , Zina Ben Miled, ID , John Wassick ² and Jeffrey Tazelaar, A Distributed Ledger for Supply Chain Physical Distribution Visibility , *Information* 2017.
- [13] Zibin Z., ,Shaoan, X., Xiangping, C., and Huaimin, W., Blockchain Challenges and Opportunities: A Survey, *Int. J. Web and Grid Services* 1, 2016
- [14] Nakamoto, S, Bitcoin: Apeer-to-Peer Electronic Cash System. Available at: <http://www.cryptovest.co.uk/resources/>, Accessed at 2017.
- [15] Tien Tuan Anh Dinh, Rui Liu, Meihui Zhang*, Untangling Blockchain: A Data Processing View of Blockchain Systems, 1041-4347 (c) 2017 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See
- [16] Krystsina S., ,Adoption of blockchain technologyu in supplychain and logistics, Bachelor of Business Logistics, 2017.
- [17] Zhen, H., Zehua, W., Wei, C., and Victor, C.M., Lenug, Blockchain-Empowered Fair Computational Resource Sharing System in the D2D Network, *Future Internet* 2017,9, 85; doi:10.3390/fi9040085
- [18] DTCC. (2016). Embracing Disruption – Tapping the Potential of Distributed Ledgers to Improve the Post-Trade Landscape, 41 Deposit Trust & Clearing Corporation. Available at: <http://www.dtcc.com/> , Accessed at. 2017.
- [19] Xiwei, X., Cesare, P., Liming, Z., Vincent, G., Alexander, P., AnBinh, T., Shiping, C., The Blockchain as a Software Connector, *Software Architecture (WICSA)*, 2016 13th Working IEEE/IFIP Conference on, 2016.
- [20] Michele, R., Floriano, S., Saverio, I., Giovanna, C., Giuseppe, L., Filippo, G., Semantic-enhanced blockchain technology for smartcities and communities, *OWL 2 Web Ontology Language Document Overview (2nd Ed.)*, W3C Recommendation, 2012
- [21] Tien Tuan, A.D., Ji, W., Gang, C., Rui, L., Beng, C.O, Kian-Lee, T. , BLOCKBENCH: A Framework for Analyzing Bitfury group ,Digital Assets on Public Blockchains, 2016 Private Blockchains, SIGMOD'17, Chicago, USA 2017.
- [22] Zibin, Z., Shaoan, X., Hongning, D., Xiangping, C., and Huaimin, W., An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends, *IEEE 6th International Congress on Big Data*, 2017.
- [23] Christopher, K. F., & Mariusz, N., From Institutions to Code: Towards Automated Generation of Smart Contracts, Foundations and Applications of Self* Systems, *IEEE International Workshops on*, 2016
- [24] Tiziana, c., From Bitcoin to The Internet of Things: The Role of The Blockchain, *Annali della facoltà giuridica dell'università di camerino – n. 6/2017*
- [25] Raja, J., & Praveen, G., Blockchain for IoT Security and Privacy: The Case Study of a Smart Home, *Conference: IEEE Percom workshop on security privacy and trust in the internet of thing*, 2017
- [26] Michael, C., Nachiappan, S.V., and Vignesh, K., Blockchain Technology, Sutardja Center for Entrepreneurship & Technology , Technical Report 2015.
- [27] Florian, K., et al., Challenges and opportunities of digital information at the intersection of Big Data Analytics and Supply Chain Management, *International Journal of Operations & Production Management*, Volume 37, Issue 1, 2015.
- [28] Michael, R., & Jan, V.B., The six core elements of business process management. In *Handbook on Business Process Management 1*. Springer, 105–122, 2015.
- [29] Asaph, A., Ariel, E., Thiago, V., and Andrew, L., MedRec: Using Blockchain for Medical Data Access and Permission Management, *2nd International Conference on Open and Big Data*, 2016.
- [30] Richard, H., Vishal, S.B., Yi-Min, C., Alin, D., Fenno, F., Terry, H., and Victor, V., Towards a Shared Ledger Business Collaboration Language Based on Data-Aware Processes, Q.Z. Sheng et al. (Eds.): *ICSOC 2016*, LNCS 9936, pp. 18–36, Springer International Publishing Switzerland, 2016.
- [31] Steve, H., Rituparna, B., Martin, W., Natalia, B., Internet of Things, Blockchain and Shared Economy Applications, *Procedia Computer Science*, Volume 98, 2016.
- [32] Marko, V., Holger, S., Oliver, S., Bernhard, M., Volker, M., Albert, M., and Tobias, K., An Approach to Optimize Data Processing in Business Processes, *VLDB Endowment*, ACM 978-1-59593-649-3/07/09, 2007.
- [33] Maria, L., & Stefanie, R., A systematic review on security in Process-Aware Information Systems – Constitution, challenges, and future directions, *Information and Software Technology* 56, pages 273–293 , 2014.
- [34] Marlon, D., Marcello, L. R., Jan, M., and Hajo, A. R., *Fundamentals of Business Process Management. Second Edition*. Springer, 2018.
- [35] Jan, v. B. & Theresa, S., Culture in business process management: a literature review. *Business Process Management Journal* 17, pages 357–378, 2011.
- [36] Ye, G. & Chen, L., Blockchain application and outlook in the banking industry. *Financial Innovation* 2, 1, 2016.
- [37] Gregor, T., Standardization in technology-based markets. *Research policy* 29, 4, pages 587–602, 2000.
- [38] Hyerim, B., Sanghyup, L., and kyeong, M., Planning of business process execution in Business Process Management environments, *Information Sciences*, Volume 268, Pages 357-369, 2014.
- [39] Gareth, W. P., & Guy, V., Overview of Emerging Blockchain Architectures and Platforms for Electronic Trading Exchanges. 2016.
- [40] Ingo, W., Xiwei, X., Regis, R., Guido, G., Alexander, P., and Jan, M., Untrusted Business Process Monitoring and Execution Using Blockchain. In *Business Process Management - 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, 2016. Proceedings (Lecture Notes in Computer Science)*, Vol. 9850. Springer, 329–347.

Defects Prediction and Prevention Approaches for Quality Software Development

Mashooque Ahmed Memon¹
Department of Computing
Faculty of Engineering, Science and
Technology (FEST) Karachi,
Pakistan

Mujeeb-Ur-Rhman Magsi
Baloch²
Department of Mathematics &
Computer Science
University of Sindh Jamshero
Hydrabad, Pakistan

Muniba Memon³,
Syed Hyder Abbas Musavi⁴
Department of Computing
Faculty of Engineering, Science and
Technology (FEST) Karachi,
Pakistan

Abstract—The demand for distributed and complex business applications in the enterprise requires error-free and high-quality application systems. Unfortunately, most of the developed software contains certain defects which cause failure of a system. Such failures are unacceptable for the development in the critical or sensitive applications. This makes the development of high quality and defect free software extremely important in software development. It is important to better understand and compute the association among software defects and its failures for the effective prediction and elimination of these defects to decline the failure and improve software quality. This paper presents a review of software defects prediction and its prevention approaches for the quality software development. It also focuses a review on the potential and constraints of those mechanisms in quality product development and maintenance.

Keywords—Software; defects; predictions; preventions; software development

I. INTRODUCTION

The software is a single entity that has a strong impact on all characteristics of software development for different domains that includes defense, medicine, science, transport, telecommunications and others. The activities of all these domain sectors constantly require high-quality software for their exact needs for the performance [1]. Software quality means being an error-free product that produces predictable results and can be delivered within a time and cost constraints [2, 3]. As a result, it very important to have appropriate approaches to develop high-quality software that can meet the increasing needs in today's business world's. The past studies suggest that no single *defect detection technology* can solve all types of defects detection problems. So, this review focuses on the effectiveness and efficiency of the defect detection process to meet the quality enhancement and cost reduction.

A “defect” is some fault or imperfection in the operation of a software product or process as a result of an error, fault, or failure. The paradigm defines the term “error” as a human action that leads to inappropriate results, and a “defect” as an erroneous decision that results in inaccurate results for a solution to a problem. A single error can result in one or more failures, and multiple failures can cause a failure. To avoid such failures in software products, defect detection activities are performed at each stage of the SDLC, depending on the needs and criticalities of the development.

Software defect identifications models [2] are very weak because they have not been able to overcome the unknown relationship between the defects and failures. The relationships understanding among them are very difficult due to the diversity of defects and failures. The “Simplified assumptions” and “heuristics” methods are frequently utilized because of the associated failures associated with failures that lead to difficult tasks for the prediction. Therefore, having an accurate defect prediction model or process in software development can able to reduce high failures and advance the eminence of the software development [4, 5]. The main cause of software failures due to its design flaws which mostly caused by the software engineers due to the misunderstanding of the requirement specifications or developing a defective code. A review study on the various domain system failure estimation suggests that 90% of the failure is due to system defects [6, 7, 21].

The approaches of defect prevention are the process for improving software quality, the core objective of that is to identify frequent causes of defects and to amend the process to avoid this kind of the defect from importunate [8]. The purpose of preventing defects is to identify them at the commencement of the life cycle and avoid them from happening again so that the defects no longer occur. Based on defect analysis, it has established to be a constructive mechanism for detecting and preventing defect requirements at the beginning phase of the software lifecycle. By analyzing the general classified defects taxonomy and past errors it can be better prevented and reliable high performing systems can be developed [11]. In terms of performance and reliability requirements, a smaller number of failures in the software requirement will affect in improved secure and quality software systems. The scope of this paper is to present an insightful exploration of the mechanisms of defect detection and defect prevention approaches that can be pursued for the quality system development processes.

The following paper presents the importance of defect prediction in Section 2 and it approaches in Section 3. In the Section 4 it presents defect prevention methodology, and Section 5 discuss its importance. Section 6 concludes and summarizes the paper.

II. IMPORTANCE OF DEFECT PREDICTIONS

In literature many empirical studies and tools [1, 5, 7, 8, 18] are designed to identify the defects for the quality software development. But these approaches can be executed at multiple points during development, not testing, which usually only happens after the executable software module is produced. A key indicates in considerate the prospective value of evaluation is that it is approximated that defects that escape from one phase of the SDLC to another, it could take an instruction for the extent to restore in the next phase. As a result, the development cost, quality, and time of the software will be significantly impacted because it is implemented at the early of the development cycle.

The software defects observed in IBM operating system depend on the field data is presented in [8], which is being classified into 408 types of defects using an "Orthogonal Defect Classification (ODC)" [16]. This classification approach is to quantify the defect, failure relation and the accuracy of prediction, 668 defects are injected over 12-open source projects. The major goal of this quantification is to show a complex relationship between software defects and failure disabilities through identifying the availability of the multiple task, such as events, conditions, etc., but the ODC approach does not allow for multiple events or conditions analysis so, user must fix it manually.

TABLE I. A SUMMARIZATION OF MERIT AND DEMERITS OF EXISISTING DEFECT PREDICTION APPROACHES

REF#	Approach	Merits	Demerits
[13]	This paper has proposed seven test effort allocation strategies utilizing the complexity measure for Fault Prediction.	<ol style="list-style-type: none"> 1. A software test simulation model based on defect prediction results for evaluating the cost-effectiveness of a test work distribution strategy. 2. The simulation model estimates the number of discoverable defects in relation to a given test resource, allocation strategy and a group of test modules for defect prediction. 3. The strategy with the best defect prediction model, test effort might be reduced by 25%, but still detected many of the defects commonly found in the test, but the company needed about 6% testing effort to collect metrics, organize data, and modeling. 	<ol style="list-style-type: none"> 1. This strategy shows the best failure prediction model but requires a high amount of test effort. 2. The results show that only the suitable test strategy with adequately high defect prediction accuracy can reduce the test workload through defect prediction.
[14]	Analysis of the Exception handling through patterns process modeling	<ol style="list-style-type: none"> 1. It shows that in many cases, there are some abstract patterns to detect the relationship between exception handling functions and the specification process. 2. Emphasis is placed on the exception handling patterns observed in process modelling over the years and described using three types' process modelling notations. 	<ol style="list-style-type: none"> 1. It has found that the exception handling pattern described here is useful for increasing the level of abstraction of the process model. It provides a way to access exception handling by providing a framework of questions.
[15]	Defect and Failure data analysis.	<ol style="list-style-type: none"> 1. This solution analyses the defect and failure data of real-system case studies. 2. Exclusively discuss the causes of software failures using other defects due to localization and distribution of defects. 3. The results show that entity faults are often reasoned for many faults spread all over the system. 	<ol style="list-style-type: none"> 1. It reveals the nature of defects and failures, and defects-defects are very beneficial.
[19]	It proposed a Specification-Based Inspection approach for the programs verification.	<ol style="list-style-type: none"> 1. Systematic and rigorous inspection methods are available to take advantage of formal specifications and analysis. 2. The purpose of this method is to utilize checks to establish if each functional solution described in the specification is correctly executed by a group of program paths to contributes certain functional aspects of the specification. 3. The results show that this method perhaps more valuable at detecting "function-related faults" than PBR but may be somewhat ineffective in detecting implementation-related faults. 	<ol style="list-style-type: none"> 1. It does not provide evaluation support for powerful features related to testing, such as reading computer instructions, managing scans, and subsequent scans for code modifications.
[20]	Utilize the machine learning classifiers based on multi-function selection techniques and implement a classification-based bug prediction method using "Naive Bayes" and "Support Vector Machine (SVM)" classifiers for bug forecasting.	<ol style="list-style-type: none"> 1. The research is generally applicable to a diversity of "feature selection techniques" based on classification-based error prediction methods. 2. Several feature selection techniques are studied, which are commonly used for classification-based defect prediction. 3. These techniques reject fewer essential features before achieving most constructive classification. The complete features utilized for training is significantly decreased below 10% of the original functionality. 4. Performance analysis of different numbers of features shows that even 1% of the original features can achieve powerful performance. 	<ol style="list-style-type: none"> 1. These techniques discard less important functions for achieving optimal classification performance. 2. A basic limitation of historically based error predictions, as there possibly recent types of errors that are not so far included in the training data.

In past years, several software technologies have been developed for the integration of state-of-the-art collection technologies that manipulate and model log-based error analysis and log data; for example, "MEADEP" [35], "NOW" [36], and "SEC" [37, 38]. However, since the log-based investigation is not supported by fully automated procedures, the processing load on most analysis loads is inadequate knowledge of the system. For example, a complex algorithm is defined for rebooting the OS in the log to identify based on sequential analysis of log messages. In addition, an error that activates multiple messages in the log causes considerable effort to use the entries for the same results of the error manifestation. Preprocessing tasks are crucial for accurate error analysis [6, 22, 27, 36].

A. Monden et al. [13] proposed a simulation model for software testing by means of defect prediction outcomes to measure the cost-effectiveness of the test assignment strategy. The proposed model assessment and resource allocation strategy, various qualified defects associated with a set of modules and defect prediction results. In a case study of the small failure prediction system recognition analysis in the telecommunications domain, the outcomes of the simulation model shows that the effective scheme is to make the test workload proportional to many failures likely in the module. Through using this strategy of the failure prediction model, the test work is reduced by 25%, while detecting defects that are usually found in the testing.

The merits and demerits of most relevant defect prediction approaches have been summarized in Table I.

III. ANALYSIS OF DEFECT PREDICTION APPROACHES

In this section, we discuss the various approaches and methods for defect prediction. Most of the approaches utilize machine learning and classification methodology to perform the prediction.

A. Defect Prediction based on Patterns

The Pattern-based detection is also based on classifiers but using a unique iterative pattern for classifying sequential data [11], software trace analysis is used for defect detection. A group of distinctive features captures a repeating sequence of actions from the program implementation trajectory that is executed first. Subsequently, the best attributes for classification are selected. Using those feature sets to train the classifier model, which will be used to identify defects. The pattern processing models allow the investigation and enhancement of processes together besides that working to coordinate multiple defects and tools to execute tasks. This kind of modeling usually focuses on the specification process, that is, how every work should execute as needed. Unfortunately, the real-world processes are rarely going well according to the need. A more comprehensive analysis of this kind of process still requires detailed information on the process model and their actions that should be taken in the event in case of failure.

B. S. Lerner et al. [14] have revealed that in numerous cases for the software defect handling, there are some abstract patterns that can detect the relationship between defect handling functions and specification procedures. As in an

"object-oriented design patterns" makes the possibility of the "development", "documentation", and "maintenance of object-oriented programs", it can be considered that process patterns can assist the enhancement and maintenance of the process models. It focuses on the defect handling patterns which have observed in process modeling for many years. They also illustrate these patterns by means of three process modeling symbols with the "UML 2.0 Activity Diagram" [17], "BPMN" and "Little-JIL" [18]. It presents an abstract construction of the pattern, in addition to examples of usage patterns. It also discusses some preliminary statistics to support the arguments that are common in these models and represent their ability to use these patterns to consider the comparative merits among the symbols.

B. Defect Prediction based on Graph Mining

The methodology of Graphics mining is based on dynamic control flow that helps identify defects that might not crash a system [34]. Its functions as a simple processing through graph nodes calls to reduce the processing overhead during execution. A graph node characterizes a function and a function call to another function which is represented by an edge. The influence of everyone edge of a node is computed based on their calling frequency. The high variation in the frequency call and changes in the node structure of the graph may be the cause of the failure. If there is a problem with the data being reassigned between the methods, it may also affect the named graph because of its functional impact.

C. Defect Prediction based on ASA

The process of "Automatic Static Analysis (ASA)" [22], [27] based prediction is primarily used for physical code analysis, which is one of the oldest traditions still practiced, but automation tools are increasingly utilized for fundamental difficulty associated with "non-observance failures", "probable memory leaks", "variable usage", etc. They occupy an essential position in the development phase because they save effort and critical re-defect leak test cycles. There are many such tools which are commonly being used as, "Findbugs", "CheckStyle", and "PMD" based on Java technology. Even though this participates as a significant function in the development cycle, it is not widely used for the defect prediction in the maintenance cycle. However, systems with compatible sources for automated static analysis can be utilized as clean aspects for excellent detection mechanisms, because the errors introduced in the executing field scan are very expensive. The maintenance cycle of the ASA prediction tool does not find many defects that may perhaps guide to the failure. Research analysis for the efficiency of ASA detection tools over the open source code represents show < 3% of failures.

S. Liu et al. [19] have presented the solution to the problems of the statistical analysis system, which are generally utilized for defect detection, and suffering due to the requirement of rigidity. It sustains a methodical and strict inspection method that takes advantage of "formal-specification analysis". The intention of the process is to describe the specification of a group of routes from every tasks base program and the route specification of the program, where the program contribute to the execution of an appropriately implemented functional environment to determine whether to

use the inspection or not. A systematic, auto-generated list of functional scenarios to obtain program paths, where each path has connected to scenarios and an inspection report generated.

C. F. Kemerer et al. [21] have studied the effects of inspection rates on software quality and studied the controller for a wide-ranging of a group of features that could influence the analysis. This data comes from the "personal software process (PSP)", performs inspections and performs development group activities. Specifically, the speed of the PSP design and code review corresponds to the preparation of the test.

J. Zhang et al. [22] has presented an enhancement to the automated static analysis which can help provide high-quality products in economic production, and they perform static analysis and check for errors and customer reports on three major sectors of the development of industrial software systems for "Nortel Networks analysis". This data shows an "automated static analysis (ASA)" for an appropriate means of detecting software errors. The automated static analysis using "Orthogonal Defective Classification" schemes is effective in identifying and mapping error probes so that subsequent software creation steps can target on more difficult, functional, and algorithmic errors. Most of the flaws that appear to be determined by automated static analysis are generated by some major type of programming error, and some of these types are likely to cause security vulnerabilities. The "Statistical analysis (SA)" outcome indicates that many automated SA errors can be effective in identifying module problems. Results analysis Static analysis tools show that it complements other error detection technologies to produce economical, high-quality software products.

D. Defect Prediction using Classifiers

A classifier based on a "clustering algorithm" and a "decision tree" or "neural network" are being utilized to recognize anomalous events of detected common incidents for the prediction [11], [12]. If a defect is found, the classifier labels the defect path to systematize the classifier. Some classification criteria generally use "NaiveBayes" and "Bagging". The Bayesian classification is a "supervised learning method" and is a "statistical method" for classification [12]. It represents a basic probability model that can capture uncertainty in a model of reason that determines the probability of a result. A recent study [7], [8], [10], [12] in this province is proposed without a secondary supervisory model to capture the regular code of behavioral probability distributions in each region to recognize incidents when they behave abnormally. This information is utilized to filter more than the labeling gives to the positioning algorithm to focus on abnormal observations.

The prediction classifiers utilizing machine learning techniques [40] are recently introduced for the defects prediction in source files. A classifier is primary trained in the defects of software development and then used to prediction if the defect vision changes it will also cause errors. A disadvantage of the existing classifier-based defect prediction technique is that it does not have enough control for actual utilization due to the various machine learning functions and the prediction time is slow.

T. Mende et al. [23] has suggested that assessing the efforts consciously can measure the accuracy of defect prediction. The traditional evaluation methods such as "recall", "precision", "Alberg chart" and "ROC curve" ignore quality assurance costs, but the action is expected to be approximately proportional to the audit or review of the module. They took advantage of the measurement to the bottom to find that the required measurement accuracy was needed for the actual test.

S. Shivaji et al. [24] has typically considers numerous attribute collection techniques for classification-based error prediction methods that use "Naive Bayes" and "Support Vector Machine (SVM)" classifiers. This technology discards less significant functions in anticipation of the most constructive classification result to be achieved. The complete functions utilized in construction is considerably decreased, often down to below 10% of the original. Both "Naive Bayes" and "SVM" through attribute selection [9] present significant improvements in comparison to the F-measure of the classification in the failure prediction and results compared to those proposed in [25].

Although many case studies on failure prediction in industry record applications [28], [29], [30] few studies have been estimated by early failure detection to reduce test effort or improve software quality. P. L. Li et al. [26] reported on ABB's experience in applying field failure prediction. Their experience is about how to decide the precise modeling method and how to evaluate the actual accuracy of predictions for several versions of the time-period. They assessed the usefulness of the forecast depends on the professional view. They identified the module as vulnerable by an expert because it identified the top four error-prone errors that identify modules in the predictive model. In addition, the module priority results have been reported by the test team to be used to reveal additional errors that are probable to reason a low error in the module. Unfortunately, there is no quantitative information on the effort to further test and the number of additional leaks needed.

IV. DEFECT PREVENTION

During software development, many defects occurred during the period of the development process. It is a defect considering that defect which is injected at the early stage of a cycle and eliminates in spite of the development process [16], [31]. Therefore, error prevention is an essential element in enhancing the excellence of software processes.

Defect prevention is a quality improvement process aimed at identifying ordinary reasons of defects and altering related processes to prevent the type of error recurrence. It also improves the eminence of software products and reduces overall costs, time and resources. This allows the project to maintain a good balance of "time", "cost" and "quality". The intention of defect prevention is to recognize defects at the inauguration of the SDLC and prevent them from reoccurring so that defects do not reappear.

A. Methodology for Defect Prevention

Defect prevention is an important activity of SDLC. Most software project teams focus on defect detection and correction. Therefore, error prevention is often an ignored

component. It is, therefore, appropriate to take steps to prevent defects from being commenced into the product at an early stage in the project. These measures are inexpensive and the total cost savings achieved by benefiting from the stage later are significantly higher than the cost of defect remediation. This saves costs and resources in the initial phases of defect analysis. The "Error injection methods" and processes facilitate knowledge of error prevention. After practicing this knowledge, quality has improved. It also improves overall productivity. The methodology for the defect prevention includes three phases as follows:

1) *Identification of the defects*: The identification of the defects can be pre-structured and designed according to the activities of specific failure defects being observed. Typically, defects can be identified in design reviews, code reviews, GUI reviews, functional and unit testing activities performed at different stages of the SDLC. In case of a defect is identified, the designed classifier classifies the defect utilizing the defined defect knowledge base. In case of having a vast defect knowledge base, it is important to analyze the failure defects through a continuous learning process to have an effective classification approach.

A model to examine software quality factors, such as a list of future defect density modules are proposed by T.Khoshgvar and E. Allen [31, 32]. The input to the model is a measure of "software complexity", such as LOC, the number, and complexity of distinctive operators. Then perform a stepwise regression to find the weight of each factor. L. C. Briand et al. [33] utilized "object-oriented metrics" to predict defect classes that might contain errors and used "PCA with logistic regression" to predict defect classes that are prone to errors. S. Morasca et al. [39] utilized a "rough set theory" and "logistic regression" to predict the possibility of the modules failure in commercial software.

2) *Analysis of the Defects*: The analysis of the defects is a continuous process for improving learning quality using defect or error data. Defect analysis generally categorized based on the process dependencies and condition process activities for the improvisation of defect identification and its possible cause for the prevention. The "Root cause analysis (RCA)" approach is an effective software defect analysis mechanism which is very useful in understanding the problems of a failure. The goal of the RCA is to recognize the root reason of defects and initiate the action of defects removal from the sources by analyzing each individual defect precisely. The qualitative analysis is inadequate only by the limitations of human investigation capabilities. This ultimately improves the quality and productivity of software organizations that provide feedback to developers.

3) *Classification of the Defects*: Defect classification can be done using common "Orthogonal Defect Classification (ODC)" techniques [16] to find defect groups and types. The ODC technology classifies defects when they occur first and when the defects are fixed. The ODC methodologies for specific technologies and some management characteristics and for each defect orthogonal can mutually exclude. These

attributes provide access to all the information that comes from the root cause, pattern, and data through a tremendous amount of data that can be analyzed. A high-quality action preparation and tracking can reduce failures and enable high levels of learning.

In case of critical and large projects, it must be deeply classified to analyze and understand defects, and in the small projects, it can be classified as defects up to the initial level of the ODC to preserve time and effort. It classifies various types of defects at diverse phases of development requirements, such as specification collection, logical design, testing, and documentation.

Defect prevention has been encountered in the past to analyze future defects and to prevent these types of occurrences including special operations. Defect prevention software processes can be applied to improve the quality of one or more phases of the SDLC. From the beginning phases of the project, to prevent defects from being presented into the product, measurements are appropriate. Even these measures are low cost, and the total cost savings achieved due to the profit at the end of the phase are quite high compared to the cost of a fixed failure. Therefore, analyzing the time needed for failures at an early stage reduces costs and resources. The defect injection method and the process can realize defect prevention knowledge. After the practice, this knowledge improves the quality. It also increases overall productivity.

B. Importance of Defect Prevention

Defect mitigation strategies exist but reflect the most cost-effective expenditures reflecting the high-level test maturity principles associated with testing efforts. To detect defect errors in the development lifecycle for implementing code specifications in your design, you should avoid errors. Therefore, test strategies can be categorized into two categories: defect detection technology and defect prevention technology.

Defect prevention during application development can save significant cost and time. It is therefore also important to decrease the number of rebuild failures resulting in cost reductions, ease of maintenance of ports and reuse. Organizations must also develop high-quality systems and provide resources to make systems reliable in less time. Determining defects increases productivity precautions and can be traced back to the fact that these defects have been injected into the lifecycle phase.

The benefits of analysis software failures and defects are well known. However, there is little-detailed research based on concrete data. M. Hamill et al. [15] analyzes the defect and failure data of a two big real-time system case studies. They specifically discuss the causes of software defects by localizing and distributing defects and using other errors. The results show that individual failures occur frequently through multiple failures in the overall system. This inspection is significant because it does not sustain multiple-use heuristics and hypothesis about the precedent. Moreover, finding and fixing errors such as software errors that result in large, complex systems is often done despite the difficult and difficult development of software development.

Due to the lack of specific domain knowledge, the new and different domain software should be developed and implemented. In many cases, the appropriate quality requirements are not initially specified. Inspection work is labor intensive and requires a high level of skill. Sometimes a well-developed quality measurement may not have been identified at design time.

No software defect detection technology can solve all the problems in error detection. Similar software reviews and tests, static analysis tools (or automated static analysis) can be used to eliminate defects before the software product is released. Inspection, prototyping, testing, and proof of correctness are several ways to identify defects. Formal inspections to identify failures in the initial phases of developing the most efficient and expensive quality assurance techniques. The adoption of several required prototypes clearly helps to overcome the perceived deficiencies. Testing is one of the least efficient techniques. It may be possible to evade detection at an early stage, which is the culprit and can be found in time. Especially the accuracy at the coding level proves to be a good detection method. Create the most accurate and economical way to build software.

V. CONCLUSION

Nowadays, intrinsic demands for software reliability are growing, and high defect tolerance systems are attracting attention. This paper has discussed several defect detection mechanisms and defect prevention mechanisms in relation to recent trends in the latest technologies. This paper presented review of the importance of defect prediction and their various approaches. Although there are several methods and technologies that are used to analyze for defect detection in a software system, but not all technologies are suitable for all systems. This paper has discussed defect prediction based on patterns, graph mining, ASA, and using the classifiers. Defect prevention methodology through defect identification, analysis and classification and its importance in reducing the system failure have also been discussed. This paper concludes that selection of defect prediction and prevention should be based on the system size and its complexity to provide a more adaptable and reliable solution for defect handling and provide high-quality software.

REFERENCES

- [1] D. Bowes, S. Counsell, T. Hall, J. Petric, T. Shippey, "Getting Defect Prediction Into Industrial Practice: the ELFF Tool", IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), pp. 44-47, 2017.
- [2] Q. Song, Y. Guo, M. Shepperd, "A Comprehensive Investigation of the Role of Imbalanced Learning for Software Defect Prediction", IEEE Transactions on Software Engineering, Pp. 1 - 1, 2018.
- [3] Z. Li, X.-Y. Jing, X. Zhu, "Progress on approaches to software defect prediction", IET Software, Vol. 12(3), Pp. 161 - 175, 2018.
- [4] A. Rahman, L. Williams, "Characterizing Defective Configuration Scripts Used for Continuous Deployment", IEEE 11th International Conference on Software Testing, Verification and Validation (ICST) Pp. 34 - 45, 2018.
- [5] S. Huda, K. Liu, M. A. razeq, A. Ibrahim, S. Alyahya, H. Al-Dossari, S. Ahmad, "An Ensemble Oversampling Model for Class Imbalance Problem in Software Defect Prediction", IEEE Access, Vol. 6, Pp. 24184 - 24195, 2018.
- [6] L. Pascarella, F. Palomba, A. Bacchelli, "Re-evaluating method-level bug prediction", IEEE 25th International Conf. on Software Analysis Evolution and Reengineering (SANER), pp. 592-601, 2018.
- [7] R. Malhotra, L. Bahl, S. Sehgal, P. Priya, "Empirical comparison of machine learning algorithms for bug prediction in open source software", International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Pp. 40 - 45, 2017.
- [8] A. Dehghan, A. Neal, K. Blincoe, J. Linaker, D. Damian, "Predicting Likelihood of Requirement Implementation within the Planned Iteration: An Empirical Study at IBM", IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), Pp. 124 - 134, 2017
- [9] X. Chen, Y. Shen, Z. Cui, X. Ju, "Applying Feature Selection to Software Defect Prediction Using Multi-objective Optimization", IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Vol. 2, Pp. 54 - 59, 2017.
- [10] M. Lanza, A. Mocchi, L. Ponzanelli, "The Tragedy of Defect Prediction, Prince of Empirical Software Engineering Research" IEEE Software, Vol. 33(6), Pp. 102 - 105, 2016.
- [11] J. H. C. Wu, Jacky Keung, "Decision support for global software development with pattern discovery", 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Pp. 182 - 185, 2016.
- [12] J. Yang, H. Qian, "Defect Prediction on Unlabeled Datasets by Using Unsupervised Clustering", IEEE 18th International Conference on High-Performance Computing and Communications, Pp. 465 - 472, 2016.
- [13] A. Monden, T Hayashi, S Shinoda, K Shirai, J Yoshida, M Barker and K Matsumoto, "Assessing the Cost-Effectiveness of Fault Prediction in Acceptance Testing", IEEE Transactions on Software Engineering, DOI-098-5589, 2013.
- [14] A. S. Lerner, S Christov, L J. Osterweil, R Bendraou, U Kannengiesser and A Wise, "Exception Handling Patterns for Process Modeling", IEEE Transactions On Software Engineering, Vol. 36, No. 2, March/April 2010.
- [15] M. Hamill and K Goseva-Popstojanova, "Common Trends in Software Fault and Failure Data" IEEE Transactions on Software Engineering, Vol. 35, No. 4, July/August 2009.
- [16] P. Tiejun, Z. Leina, F. Chengbin, "Defect Tracing System Based on Orthogonal Defect Classification", International Conference on Computer Science and Software Engineering, Vol. 2, Pp. 574 - 577, 2008.
- [17] OMG, "Unified Modelling Language", Superstructure Specification, Version 2.1.1, <http://www.omg.org/spec/UML/2.1.1/Superstructure/PDF/>, 2010.
- [18] A. Wise, "Little-JIL 1.5 Language Report", technical report, Dept. of Computer Science, Univ. of Massachusetts, 2006.
- [19] S. Liu, Y Chen, F Nagoya and J A. McDermid, "Formal Specification-Based Inspection for Verification of Programs", IEEE Transactions on software engineering, vol. 38, no. 5, 2012.
- [20] Y. Kamei, S. Matsumoto, A. Monden, K. Matsumoto, B. Adams, and A. E. Hassan, "Revisiting common bug prediction findings using effort aware models", Proc. 26th IEEE Int'l Conference on Software Maintenance (ICSM2010), pp. 1-10, 2010.
- [21] C. F. Kemerer and Mark C. Paulk, "The Impact of Design and Code Reviews on Software Quality: An Empirical Study Based on PSP Data", IEEE Transactions on Software Engineering, Vol. 35, No. 4, July/August 2009.
- [22] J. Zheng, L Williams, N Nagappan, W Snipes, J P. Hudepohl and M A. Vouk, "On the Value of Static Analysis for Fault Detection in Software", IEEE Transactions on Software Engineering, Vol. 32, No. 4, April 2006.
- [23] T. Mende and R. Koschke, "Revisiting the evaluation of defect prediction models", Proc. Int'l Conference on Predictor Models in Software Engineering (PROMISE'09), pp. 1-10, 2009.
- [24] S. Shivaji, E. J Whitehead Jr., R Akella and S Kim, "Reducing Features to Improve Code Change-Based Bug Prediction", IEEE Transactions on Software Engineering, Vol. 39, No. 4, April-2013.
- [25] S. Kim, E. Whitehead Jr., and Y. Zhang, "Classifying Software Changes: Clean or Buggy?" IEEE Trans. Software Eng., vol. 34, no. 2, pp. 181-196, Mar./Apr. 2008.

- [26] P. L. Li, J. Herbsleb, M. Shaw, and B. Robinson, "Experiences and results from initiating field defect prediction and product test prioritization efforts at ABB Inc.", Proc. 28th Int'l Conf. on Software Engineering, pp. 413-422, 2006.
- [27] F. Wedyan, D. Alrmony, and J. M. Bieman, "The Effectiveness of Automated Static Analysis Tools for Fault Detection and Refactoring Prediction", ICST '09. International Conf., Vol. 1(4), pp.141-150, 2009.
- [28] N. Ohlsson, and H. Alberg, "Predicting fault-prone software modules in telephone switches", IEEE Trans. Software Engineering, vol. 22, no. 12, pp. 886-894, 1996.
- [29] T. J. Ostrand, E. J. Weyuker, and R. M. Bell, "Predicting the location and number of faults in large software systems", IEEE Trans. on Software Engineering, vol. 31, no. 4, pp. 340-355, 2005.
- [30] A. Tosun, B. Turhan, and A. Bener, "Practical considerations in deploying AI for defect prediction: a case study within the Turkish telecommunication industry", Proc. 5th Int'l Conf. on Predictor Models in Software Engineering (PROMISE'09), pp. 1-9, 2009.
- [31] T. Khoshgoftaar and E. Allen, "Predicting the Order of FaultProne Modules in Legacy Software", Proc. Int'l Symp. Software Reliability Eng., pp. 344-353, 1998.
- [32] T. Khoshgoftaar and E. Allen, "Ordering Fault-Prone Software Modules", Software Quality J., vol. 11, no. 1, pp. 19-37, 2003.
- [33] L. C. Briand, J. Wiist, S.V. Ikononovski, and H. Lounis, "Investigating Quality Factors in Object-Oriented Designs: An Industrial Case Study", Proc. Int'l Conf. Software Eng., pp. 345-354, 1999.
- [34] A. A. S. Haghghi, M. A. Dezfuli and S. M. Fakhrahmad, "Applying mining schemes to software fault prediction: A proposed approach aimed at test cost reduction", In Proceedings of the World Congress on Engineering, pp.415-419, 2012.
- [35] D. Tang, M. Hecht, J. Miller, and J. Handal, "Meadep: A Dependability Evaluation Tool for Engineers", IEEE Trans. Reliability, vol. 47, no. 4, pp. 443-450, Dec. 1998.
- [36] A. Thakur and R.K. Iyer, "Analyze-Now-An Environment for Collection and Analysis of Failures in a Networked of Workstations", IEEE Trans. Reliability, vol. 45, no. 4, pp. 561-570, Dec. 1996.
- [37] R. Vaarandi, "SEC-A Lightweight Event Correlation Tool", Proc. Workshop IP Operations and Management, 2002.
- [38] J. P. Rouillard, "Real-Time Log File Analysis Using the Simple Event Correlator (SEC)", Proc. USENIX Systems Administration Conf., 2004.
- [39] S. Morasca and G. Ruhe, "A Hybrid Approach to Analyze Empirical Software Engineering Data and Its Application to Predict Module Fault-Proneness in Maintenance", J. Systems Software, vol. 53, no. 3, pp. 225-237, 2000.
- [40] V. Challagulla, F. Bastani, I. Yen, and R. Paul, "Empirical Assessment of Machine Learning Based Software Defect Prediction Techniques", Proc. IEEE 10th Int'l Workshop Object-Oriented Real-Time Dependable Systems, pp. 263-270, 2005.

Design and Implementation of a Risk Management Tool: A Case Study of the Moodle Platform

Nadia Chafiq

Multidisciplinary Laboratory in Sciences and Information,
Communication, and Educational Technology (LAPSTICE),
Observatory of Research in Didactics and University
Pedagogy (ORDIPU)
Faculty of Sciences Ben M'Sik, Hassan II University of
Casablanca, B.P 7955 Sidi Othmane, Casablanca, Morocco

Mohammed Talbi

Observatory of Research in Didactics and University
Pedagogy (ORDIPU)
Faculty of Sciences Ben M'Sik, Hassan II University of
Casablanca, B.P 7955 Sidi Othmane, Casablanca, Morocco

Mohamed Ghazouani

Systems architecture team - ENSEM- Casablanca, Morocco

Abstract—During the last years, the distinctive feature of our society has been the rapid pace of technological change. In the Moroccan context, universities have put digital learning at the heart of their projects of development thanks to a wide range of hybrid training devices, Small Private Online Course (SPOC) and Massive Open Online Courses (MOOCs) via Virtual Work Environment (ENT, Environnement Numérique de Travail). On the one hand, the purpose of using these devices consist in helping improve their performance and in enhancing their attractiveness. On the other hand, is aimed at meeting the increasingly diverse student's needs, thanks to the infrastructures reorganization and a renovated pedagogy. Also, extensive use of information and communication technologies at different universities exposes them to a problem related to information system (IS) risks in general and e-learning in particular. The risk assessment is quite complicated and multidimensional. It must take into account many components, including assets, threats, vulnerabilities, controls already in place and analyses. In this work, we first propose the methods of risk management. We then present the risk analysis related to the Moodle platform.

Keywords—Risk management; e-learning; mehari; platform

I. INTRODUCTION

The universities of today have at their disposal an exceptional potential to exploit: collaborative platforms, the blockchain, deep learning, serious games, rapid learning, virtual classes/video conferencing, mobile learning, MOOCs all tools accessible and navigable from all devices with improved ergonomics [1]. These new practices also concern Big Data, learning analytics and performance management. However, the rapid evolution of these technologies is increasing the risks related to digital learning in particular and information systems in general (risks related to the storage and transmission of data, etc.) [2]. The risk is an integral part of the management of a digital project. Hence, it is essential to have a risk management plan of at an early stage of the projects. Risk management then becomes a strategic function, an integral part of the university's operational and strategic management process. Thus, IS risk management defined as a mechanism for

identifying and analyzing risks to information systems, to determine security objectives and implement security measures to achieve these objectives. However, the absence of an IS risk management strategy favours the appearance of many facets of risk. Currently, several standards and methods are available internationally are working to sustain a high level of protection and performance for the IS. Risk management can be applied across the organization, in all areas and at all levels, at any time, as well as to specific functions, projects and activities. So, what are the risks related to the digital device? Moreover, what are the standards and IT standards used to manage risks related to e-learning? In the first part of this article, we discuss the methods of risk management. The second part is devoted to a case study, and in the third part, we present an application dedicated to risk management about online platforms.

II. AVAILABLE RISK MANAGEMENT METHODS AND TOOLS

Currently, universities are evolving in a complex, uncertain and changing environment. The Moroccan university, for instance, Hassan II University of Casablanca (UH2C), should face the emergence of more and more diversified risks. For this reason, the UH2C has set up, since 2013, a Directorate General of Information Systems Security (DGISS). Faced with this environment of "less and less predictable," it is becoming increasingly urgent for all institutions to put in place a risk management system that will identify, assess and manage both actual and potential risks. Several methods of risk analysis are currently available, and the primary concern of decision-makers is to choose the most appropriate method in the context of their organizations. This section provides an overview of existing risk management methodologies and tools. Table I lists the main well-known methods and associated tools.

Before investing in one method or another, it is essential that the chosen method meet the requirements of the organization. Also, a risk management method is an analytical tool for identifying risks and by proposing solutions to address it. Risk management methods are based on different analysis strategies. Methods exist for covering different perspectives in risk management, for example, EBIOS, MEHARI and

OCTAVE. It should note that, while there are concepts familiar to all these methods, each method has a different way of performing an information system risk assessment. Among these tools, we propose an integrated use of ISO27005 and Mehari for the implementation of an e-learning risk management platform (PGRE, platform de Gestion des Risques en e-learning) (see Fig. 1).

TABLE I. RISK MANAGEMENT METHODOLOGIES AND TOOLS AVAILABLE

Available risk management methodologies			
Au IT Security Handbook	Cramm	A&K Analysis	Ebios
ISAMM	ISF Methods	SP800 30 (NIST)	ISO/IEC 27005
Mehari	MIGRA	Octave	Risksafe Assessment
Available risk management tools			
Gstool	Cramm	MetricStream	Ebios
Callio	Riskwatch	ISAMM	Mehari
vsRisk	Risicare	CCS Risk Manager	Cobra
MIGRA Tool	Modulo Risk Manager	Proteus	Octave



Fig. 1. PGRE Platform.

The reason for ISO 27005 is to give rules to information security risk management. It bolsters the general concepts indicated in ISO 27001 and is intended to assist the satisfactory implementation of information security based on a risk management approach. [3]. It does not indicate, prescribe or even name any specific risk analysis method, although it specifies a structured, systematic and rigorous process from analyzing risks to creating the risk treatment plan [4]. MEHARI is agreeable to ISO 13335 risk management standard. It allows the stakeholder to develop security plans, based on a list of vulnerability control points and an accurate monitoring process to achieve a continual improvement cycle.

III. CASE STUDY: RISK MANAGEMENT OF THE MOODLE PLATFORM

The information system of the UH2C is composed of several applications, for example, APOGEE: Application for the management of students and teaching mainly the administrative and pedagogical management of schooling, a platform dedicated to MOOCs and Moodle: e-learning platform intended for distance education. For our case study, we will focus on the study of risks related to the latest

application, namely the Moodle platform. The educational platform of the UH2C is accessible from the ENT (see Fig. 2).



Fig. 2. Moodle Platform-UH2C.

The educational platform (Moodle) is a teaching/learning environment. It is made available to teachers to enrich and accompany classroom teaching. This study, therefore, concerns the analysis of the risks that such a platform may experience when used in the service of distance education.



Fig. 3. Risks Related to the Moodle.

The risk assessment is quite complicated and multi-dimensional. It is a question of categorizing the goods, the processes and the activities of the organization, to identify the perimeters of the risks, to define the risks and to establish the typology of the latter for example, there are the economic risks and financial (hosting), environmental risks Energy consumption (data center), technical risks (data loss, migration of data from existing courses on a platform to a new platform). Legal risks (copyright, disclosure, legislation and regulations) and Risks related to new pedagogical approaches integrating new technologies (not available online tutors/neglect of the interactivity aspect of the learning process). To evaluate the risks related to the Moodle platform (see Fig. 3), the auditor of the information system aims to:

- constitute a database
- inventory all identified risks
- evaluate the criticality of these risks (gravities and frequency)
- propose corrective actions
- define the aspects to be strengthened about the control structures (organization, attributions and functioning of these entities, training and competent human resources, methods and tools for work)
- propose an action plan and audits to programme in the next five years.

These steps have integrated into the platform (PGRE) (see Fig. 4).



Fig. 4. Risk Assessment Process at the PGRE Platform.

The application executes an input questionnaire, which utilized for asset impact evaluation. Risk values are computed, and in light of threats, assets and risks, suitable measures proposed by the system. By these attributes, the manager can choose to implement them or not. At last, the system produces a study summary report and an action plan suggesting the manager countermeasures to implement. This section discusses its components.

A. Risk Identification

Risk identification is a process that can necessarily be done from a knowledge base. MEHARI proposes a knowledge base of risk scenarios that can be used by the vast majority of organizations. Nevertheless, it is possible to develop variants, to complete this base, or to develop new ones, relying on a specific guide. We will then work on the MEHARI 2010 knowledge base. The MEHARI 2010 knowledge base contains nearly 800 standard risk scenarios [5]. Of all these scenarios, some may be genuinely critical and deserve detailed consideration, while others may not be relevant to the entity or deserve attention. It may, therefore, be considered desirable to make a selection of scenarios before addressing a detailed assessment of their severity and a risk treatment plan.

We will present in the following paragraph some risk categories related to the Moodle platform. The significant threats are as follows :

- **Economic and financial risks:** Hosting internally inducing very high costs.
- **Environmental risks:** excessive storage volume increase, in the absence of an outsourcing policy, leading to an increase in energy consumption by adding database servers.
- **Technical risks:**

Inoperative features: Business interruption of local network services, due to a long-term absence of (internal) staff.

The hijacking of application data files in operation, by an unauthorized third party, connecting from outside to the local network.

The obligation to leave the platform for further investigation, through the search engines, to locate the titles of the appropriate short.

- **Legal risks:** Non-compliance with legislation or regulations relating to the protection of intellectual property due to non-application of procedures, by lack of knowledge.
- **Risks related to new pedagogical approaches integrating new technologies:**
Risks of business as usual (Neglect of the interactivity aspect of the learning process).
Not available online tutors (Keeping online tutors unsatisfied) [7].

B. Risk estimation (Impact * probability)

The manager must select the threats and indicate its frequency. The frequency of the threat is never exact. The manager should be based on specific information such as attacks and incidents detected on the threat that the organization is facing. By using these parameters, the manager can provide a rough estimate of the frequency of a particular threat. Risk Estimation handles the execution of the impact and the probability calculation.

C. Risk Evaluation

Risk Evaluation has the part of characterizing the risk given the ISO27005 risk assessment matrix. This part is to classify risk levels according to different levels of gravity. Using this matrix, the manager has an on-screen overview of all risks and their classifications.

D. Risk Treatment

Risk Treatment presents all the threats to each asset. Each line of the application also contains the level of risk and a drop-down menu offering the following options [6]: attenuate, transfer, accept and avoid :

- **Risk mitigation:** If the manager chose to mitigate the risk, the system suggests administrative controls, technical or physical to be applied within the Moodle platform according to their effectiveness and cost of implementation.
- **Risk avoidance:** Decide to avoid the risk by eliminating the risk situation by structural measures.
- **Risk acceptance:** the manager accepts the risk as it is.
- **For any risk accepted or reduced causing residual risks** after the traditional preventive measures will be transferred after that to other organizations able to better manage these risks (Risk transfer).

Users of the Moodle platform of UH2C face different risks or threats, as indicated in the paragraph above. The following measures are proposed by the application (PGRE) to minimize these risks:

- **Remedies of economic and financial risks:** Outsourcing hosting
- **Remedies of environmental risks:** the implementation of ventilation devices and more efficient cooling systems the server room must be well ventilated; Free cooling of Datacenter.

- **Remedies of technical risks:** Replacement by a resource person whose presence is sustainable ; Install debugger programs to help programmers detect bugs and send them later ; Reduce the size of the Moodle platform exposed to hackers by adding a proxy server upstream ; Raise awareness about the use of internal search engines.
- **Remedies of legal risks:** User awareness of current legislation at the start of registration and all stages of training.
- **Remedies of risks related to new pedagogical approaches integrating new technologies:** Pedagogical re-engineering is taking into account the aspect of interactivity internally.

IV. ANALYSIS AND DISCUSSION

The contribution of our work consists in proposing a risk management tool PGRE in e-learning (PGRE, the platform of risk management in e-learning) adapted to university establishment based on international methods and standards. This tool covers the entire methodology of risk analysis from assessment to risk management. In this article, we have detailed the components of the PGRE platform.

The experimentation with the tool constituted the first phase of the deployment of our platform. We did the first experiment via the Moodle application. Also, this phase of experimentation allowed us to realize the modalities of the concrete use of the tool and to reveal the contributions and the limits. Also, the experimentation of the PGRE platform has encountered many difficulties:

Forgetting a risk: One of the issues of recurring concern is the fear of having forgotten a risk or of having made an error of appreciation in the analysis. The causes of error are many: lack of a critical element in the inventory of assets related to e-learning, the problem of identification of threats and error in the valuation of an asset.

Lack of monitoring of the implementation of security measures: Experience shows that once a project is completed, the risk assessment that has been carried out during the project is classified, with no updating work planned.

Lack of communication: Communication in the field of risk management in academic institutions remains delicate and is often hampered by a number of problems : Communication is often hampered by the geographical dispersion of university campuses (for example the University Hassan II of Casablanca (UH2C) contains 18 educational establishments spread over 6 university campuses) and the breakdown of the interveners within the directorates or services that can hinder or prevent the holding meetings.

In perspective, we are continuing our work to finalize the experimental PGRE platform, adding a measurement evaluation module after their implementation and follow-up.

ACKNOWLEDGMENT

I would like to thank my advisor Mr. Talbi, Ph.D., for their invaluable guidance and many useful suggestions during my work on this paper. I would also like to express my gratitude to all those who gave me the possibility to complete this paper.

V. CONCLUSION

In this paper, we introduced the application components (PGRE) to manage the risks of the Moodle platform. This particularity of our approach is that integrated use of ISO27005 and Mehari to design a comprehensive Information Security Risk Management Tool. The system applied to a concrete example (case of the Moodle platform of Hassan II University of Casablanca). Future work consists in testing the risk management platform related to e-learning (PGRE) on other distance learning devices, for example, MOOCs based on the Open EDx platform, SPOC to overcome obstacles and achieve the goals of distance education.

REFERENCES

- [1] CHAFIQ, N., BENABID, A., BERGADI, M., TOURI.B., TALBI, M., LIMA, L.: Advantages and Limits of the Implementation of Blended Learning for Development of Language Skills in Scientific Students, *Procedia - Social and Behavioral Sciences* 116, 1546-1550. Page 1549, 2014
- [2] Chafiq, N., Benabid, A., Bergadi, M., Lima, L.: Intérêts et limites de la mise en oeuvre d'un dispositif hybride pour le développement de la compétence langagière chez les étudiants scientifiques, *Le langage et l'homme: Revue de didactique du français*, ISSN 0458-7251, Vol. 47, N. 1, pgs. 111-119, 2012
- [3] Information technology—Security techniques— Information security risk management. INTERNATIONAL STANDARD ISO/IEC 27005 First edition 2008.
- [4] Mohamed GHAZOUANI, Hicham MEDROMI, Brahim BOULAFDOUR and Adil SAYOUTI, "A model for an Information security management system (ISMS Tool) based multi-agent system." International Conference on Intelligent Information and Network Technology (IC2INT'13)
- [5] GHAZOUANI, Mohamed, MEDROMI, Hicham, SAYOUTI, Adil, et al. Article: An Integrated use of ISO27005, Mehari and Multi-Agents System in order to Design a Comprehensive Information Security Risk Management Tool. *International Journal of Applied*, vol. 7, p. 10-15.
- [6] GHAZOUANI, Mohamed, FARIS, Sophia, MEDROMI, Hicham, et al. Information Security Risk Assessment--A Practical Approach with a Mathematical Formulation of Risk. *International Journal of Computer Applications*, 2014, vol. 103, no 8. <https://pdfs.semanticscholar.org/1c40/467699b011318a49d21191c272228b94dc1d.pdf>
- [7] Chafiq, N., Talbi, M., Tutoring Functions in a Blended Learning System: Case of Specialized French Teaching, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 2, page 25, 2017

Artificial Neural Network based Weather Prediction using Back Propagation Technique

Saboor Ahmad Kakar¹,
Naveed Sheikh²

Department of Mathematics
University of Balochistan Quetta

Saleem Iqbal⁴, Abdul Rehman⁵,
Aziz ullah Kakar⁶

Department of Mathematics
University of Balochistan Quetta

Hazrat Ali Kakar⁸
Department of Mathematics
University of Balochistan
Quetta

Adnan Naseem³

Department of Computer Science
COMSATS University Islamabad

Bilal Ahmad Kakar⁷

Department of Computer Science
University of Balochistan Quetta

Bilal Khan⁹

Department of Computer Science
COMSATS University Islamabad

Abstract—Weather forecasting is a natural phenomenon which has some chaotic changes happening with the passage of time. It has become an essential topic of research due to some abrupt scenarios of weather. As the data of forecast is nonlinear and follows some irregular trends and patterns, there are many traditional techniques (the literature like nonlinear statistics) to work on the efficiency of models to make prediction better than previous models. However, Artificial Neural Network (ANN) has so far evolved out to be as a better way to improve the accuracy and reliability. The ANN is one of the most fastest growing technique of machine learning considered as non-linear predictive models to perform classification and prediction weather forecasts maximum temperature for the whole days (365) of the year. Therefore, a multi-layered neural network is designed and trained with the existing dataset and obtained a relationship between the existing non-linear parameters of weather. Eleven weather features were used to perform classification of weather into four types. Furthermore, twenty training examples from 1997-2015 were used to predict eleven weather features. The results revealed that by increasing the number of hidden layers, the trained neural network can classify and predict the weather variables with less error.

Keywords—Weather forecasting; artificial neural network; classification; prediction; backpropagation; hidden layers

I. INTRODUCTION

Due to sudden changes happens in the nature and weather conditions, the massive amount of power is always required to manage and keep the balance for the atmosphere. Weather forecasts the natural phenomena, involved the vague understanding of atmosphere processes due to the time difference. Weather forecast is a process of rapidly changing conditions like humidity, rainfall, wind speed, directions, temperature, etc. It is used to collect this data with some defined tools like, meteorological satellites, radars, wireless sensor and high-speed computers to perform computation. The data which is recorded through satellites are always available in the form of an images to get the whole idea of abrupt changes happened in the environment. Apparently, weather is an intensive, chaotic and dynamic process makes it a fierce challenge.

As whole it is considered, as an important matter to be focused on all over the world, because the weather prediction can have expressive impact on variant parts of the world. This prediction activity is particularly handled by individuals to improve the efficiency of operations performed on daily basis. It is one of the most essential and in demand responsibility taking over by meteorological services. And all the decisions are taken under the uncertain situation associated with local and global climate variables. Different researchers and scientists are performing research since 1920's to predict the disasters and changes in climate conditions. The most important factor that varies is time series. It involves a combination of observation, trends and patterns both for regional and global. So far, researchers are trying to predict the weather forecasting by different means which increase the computational power. As we are talking about prediction which means to determine something going to happen in future. In a past few decades, the excessive increment in a computational power has forced to design and develop the complex systems having the ability to deal with the hardness of power systems to perform calculations. Machine learning and data mining [13, 14, 16] are commonly used fields for this research as these are actively used to improve the computational power. For weather forecast most of the systems depends on conditions to adjust to prevent the loss or damage to some extent.

The weather forecast systems need to be intelligent that they can easily read the statistical data to generate patterns and rules to study and based on past data predict the future. Recent study [15] has shown that numerical weather prediction model could be introduced to represent the global importance of forecast. They simulated the sub grid lakes in global forecasts for reduction at high latitude in forecast errors especially in spring and summer. Although, above described tools like radars, wireless sensors and satellites are more enough for handling conditions of weather but these improvements and enhancement are result of betterment for future time. And the better observation under computational techniques. Since, many researchers have made efforts to identify the well-defined forecasting models which includes linear, non-linear and hybrid techniques to improve accuracy. It has been

observed from recent studies that in some cases the hybrid computational models [17,18] results in decomposition of linear and nonlinear forms and considered as efficient approach comparative to single or individual techniques. It also has less number of chances to produce errors. On contrary some of studies claims that hybrid or combination of models not proved to be efficient. Such kind of uncertainties in this case always open the door for new opportunities to make models more precise. Even due to high penalty of imbalance systems relying always on meteorological forecast is not a good option.

This research is focused to propose an efficient model based on ANN to improve the computational power to make future prediction more efficient. The selected features to perform the experiments for actual weather is sunny, foggy, thunderstorm and rainy for the course period of 20 years (7000 days).

The study is categorized as: section II represents the study done previously, section III show the methodology has been used for this research, section IV demonstrated the results along with discussion and at last conclusion of the study is presented.

II. LITERATURE REVIEW

Most of the work have been done on different weather forecasting applications by using prediction techniques especially in machine learning field. Some of the work is summarized in this section.

Authors in [1] focused on reduction of random and numerical errors with the post processing of Numerical weather predictions (NWP) models to propose the best error free model which will be used for weather forecasts purposes. They had used most inspiring the Kalman filter to perform analog forecast in an ordered form. Along with it, adapted the weighing strategy to verify the forecast analog for specific time and location. With this new approach, they have successfully achieved the improvement in comparison of other algorithms. With the strong combination of both, NWP altogether and analog by producing the hybrid of these two models with logistic regression as a preprocessing [2]. Authors had cover up the probabilistically 2 m weather forecast with their efficient hybrid technique as shown to outperform comparatively others. Somehow this is not performing good for 10 m speed (wind). They considered, the representative NWP for long term purpose to overcome the wind speed factors involved like ambient temperature, atmospheric pressure, local terrain, etc. [3] Authors included the machine learning by joining the neural networks with NWP to assess the wind speed which offered a satisfying improvement in performance. They have proposed the new method with the use of feed forward artificial neural network for weather prediction which is proved to be a better approach as compared to traditional approaches [4]. Different datamining techniques has also been used for classification of weather parameters such as, the C5 decision tree with the consideration of some basic factors like min and max temperature, wind speed and rainfall prediction for a specific month or year [5]. These techniques been observed in the months by giving enough data to influence the weather prediction with some expected changes in identified climate patterns. Authors used a combination of TLFN neural network

and RNN architectures to make ANN models through observe pattern variables to detect the relationships for future predictions. At the same time, their prediction analysis was performed with linear regression to achieve the better results of prediction with great number of accuracy [8]. Multi classification is proved to be great in their research by considering the large number of features. According to given results, their model is particularly providing an efficient prediction. The classification techniques are a possible solution of severe occurrences like socio economic or environmental [6]. However, the large number of meteorological data makes it difficult or somehow impossible to perform analysis for weather prediction. By considering the classification approach of ANN we can conclude that their approach is far better than traditional to identify associated patterns for consecutive events.

Consecutively, there are more complex strategies to leverage the load forecasting by using hierarchy patterns in distributions network for small area [7]. It is used for parameter optimization to make innovative forecast. Authors proposed a new approach named Dilation erosion perceptron with a fusion of modified genetic algorithm which is based on a complete theory of mathematical morphology for weather prediction [9]. They have also considered to remove the temporal distortions in weather forecasting by proposing a method the automatic phase fix procedure. And their evolutionary process model demonstrated the good results. In [10] somehow, we find it difficult to figure out the prediction patterns because of some consistencies happened in characteristics of weather. Authors targeted the occurrence of changes specifically conducting in a huge region of Bangladesh. Their model shows reliable prediction for this seasonal change. Classification is commonly encountered as a more powerful decision-making idea in ANN [11]. Even appeared as an evolutionary task in data mining and machine learning fields. Neural Networks are being applied for different mathematical and optimization purposes such as forecasting, image processing, object recognition, classification, prediction, processing of signals, etc. Authors considered an important factor to perform some operations for solar online power forecasting, with the selection of advanced statistical methods. Their online solar power application got trained for operational planning of system. One of the benefit could be noticed of ANN based photovoltaic system is that it could handle the abrupt circumstances or variations by this system. Proposed radial based function network makes prediction for 24 hours. They have proved that given model worked with 90% reliability for crystal clear sky on the other hand 80% for unclear sky condition as the weather uncertainties always led to great errors. In this research, the authors proposed two ANN based approached one of them is used d to predict the rainfall rate for 2001 to 2013 period through sensitivity analysis [12]. However only one model is proved to be efficient as compared to other one.

III. METHODOLOGY

This section introduces the methodology being used and the analysis performed. In this research, basically two things are considered; weather classification and weather prediction. Further, the weather data is classified into four types thunderstorm, Rainy, Foggy, and sunny then predicts weather

features for next year after training on 20 years on ANN. And the 20 training examples are the data acquired for the year 1997 – 2015. There are 11 prediction features have been taken into account which are MaxDewpoint, MaxHumidity, MaxPressure, MaxTemperatue, MaxVisibility, MeanWindSpeed, MinDewpoint, MinHumidity, MinPressure, MinTemperatue, MinVisibility.

A. Artificial Neural Networks

ANN has been used since 1943 and inspired from human brain reactions. Just like human brain needs several neurons to transmit information and perform the actions just like that ANN needs neurons to perform the action in hidden layers. The general architectural diagram of ANN has been shown in Fig. 1. They are capable enough to train or learn more complex functions. It consists of input layer, hidden layer and output layer. The number of neurons is equal to the number of input variables present. The complexity of model always depends on the scenario to perform action iteratively. If you want your model to perform better and efficient activities, you need to pass it to large number of data and number of hidden layers. Then some activation functions like hyperbolic, tangent, sigmoid, SoftMax, etc. makes the work easier to accommodate the model they evaluate the values passed to it to compute error value. It has multiple types of neural networks. A neural network which is made up of processing units has a natural ability to store knowledge. Knowledge in network is acquired from the environment through given learning process. And the weights are the strengths of neurons to store the knowledge.

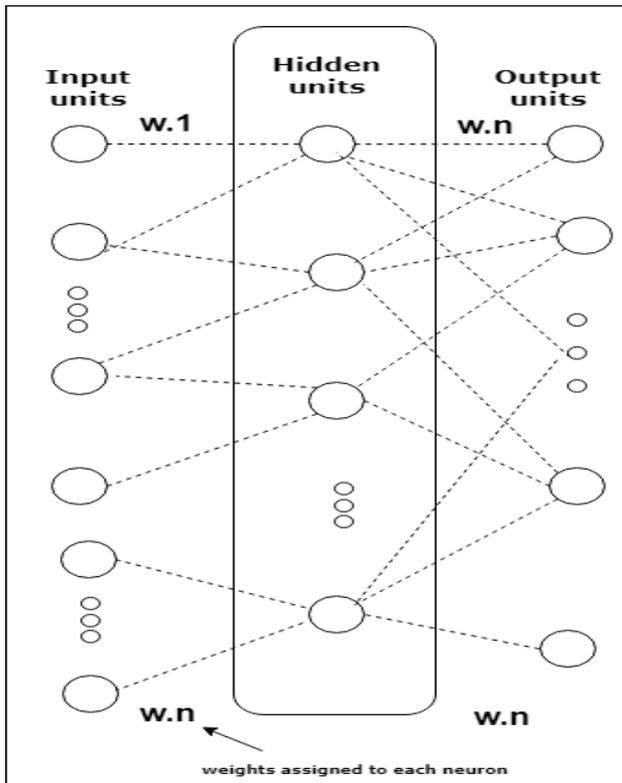


Fig. 1. General Architecture of ANN.

B. Architecture of ANN based Model

Our proposed architectural model is shown in Fig. 2. This diagram represents that increased number of neurons in hidden layers always raise the ability of network to perform efficiently by getting data from input nodes. It is also called as fully connected network; if every single node of one layer is connected to every other node in adjacent layer. The theta values in the network represents the weights as they stores the processing information of neurons attached with the layers. After training the data the network hidden layers conclude the desired output for a set of previously unseen input data. The model is closely prepared to process the information through neurons. The data is gathered from what the project predicted and what the actual weather was. The mathematically general model is explained step by step below. The parameters have been used in the algorithm are:

- backPropClassifier – has the classification algorithm through back propagation
- backPropPredictor – has the prediction algorithm through back propagation
- forwardPropClassifier – has the classification algorithm through forward propagation
- forwardPropPredictor – has the prediction algorithm through forward propagation
- Classifier – has the main classification algorithm.

- 1) Load a file and save it in a variable 'X'. All files of weather data are gathered from the period '1997-2015' are loaded.
- 2) Number of features (total 11 features+5 bias units) = Number of features (Max humidity, Mean wind speed, etc.) + bias units.
- 3) $Y=X(:, \text{Number of features}) \rightarrow$ classes of weather are extracted from X; such as ':' represents 16 column in Y.
- 4) $X(:, \text{Number of features}) = [] \rightarrow$ 16 column of X is emptied.
- 5) Weather classes in Y are bit mapped \rightarrow they are bit mapped is because it is easy to match numerical output.
- 6) This is how we get [X Y] array. such as X is an input and Y is output.
- 7) Now we will define some variables: -
 - IEPSILON = 1 (weight) \rightarrow Initial theta of each layer
 - LAMBDA = 0.01 \rightarrow Regulation parameter
 - NUMBER OF FEATURES \rightarrow Number of weather features
 - M \rightarrow Number of training examples
 - K \rightarrow Number of output classification
 - m = 365 \rightarrow DAYS IN A YEAR

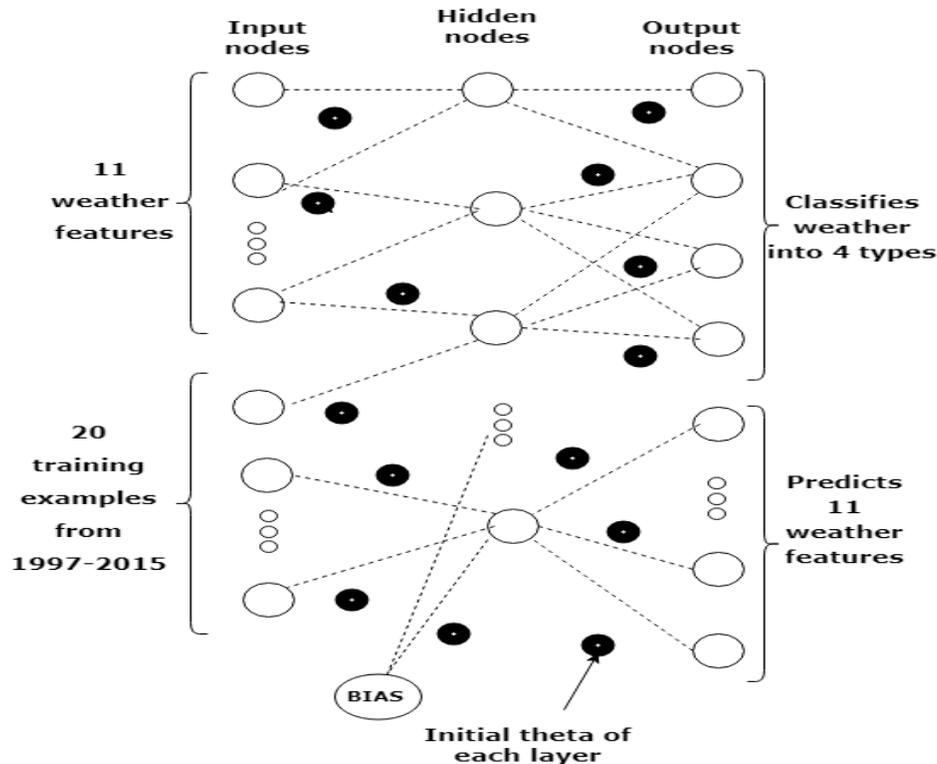


Fig. 2. Proposed Model of ANN Architectural Diagram.

8) Then we choose initial theta matrix for layer 1 and layer 2 of Neural Network through formula: -

- $I\theta_{1} = 2 * I\epsilon * \text{rand}(\text{Number of features}-1, \text{Number of features}) - I\epsilon * \text{ones}(\text{Number of features} - 1, \text{Number of features})$

9) 'Predictor' is called to train Neural Network:

- It takes array X, $I\theta_{1}$, $I\theta_{2}$, LAMBDA as input.
- Few variables are already defined assigned to new ones: -
 - $E = M-m;$
 - $x = X(:, 1)$
 - $y = X(:, 366)$
- A loop is executed to separate 365 values of a single feature.
- $I\theta_{1}$ & $I\theta_{2}$ are placed inside thetavec.
- Lost function is minimized through 'prediction cost function'
 - Optimset \rightarrow create structures for optimization functions
 - Fminuc \rightarrow tries to minimize a function.

▪ Steps:

- 1) Reshape thetavec & get $I\theta_{1}$ and $I\theta_{2}$ from it.
 - 2) Apply forward propagation:
 1. Input x, θ_{1} , θ_{2}
 2. output a1, a2, a3
 - 3) Computing error
 1. $\text{sqrError} = \text{sum}(\text{sum}((a3-y).^2))$
 2. $\text{sqrError} = \sum(\sum(a3-y)^2)$
 - 4) Computing cost function
 1. $Jval = (1/(2*e)) * (\text{sqrError} + \text{LAMBDA} * \text{sum}(\text{thetavec.*2}))$
 2. $Jval = 1/2e(\text{sqrError} + \text{LAMBDA} * (\sum \text{thetavec}^2))$
 - 5) Apply backpropagation through 'backpropPredictor'
 1. Input \rightarrow a1, a2, a3, θ_{1} , θ_{2} , y, e, lambda
 2. Output \rightarrow [gradient vector]
 - Reshape optimal theta vector into matrices for layer 1 and layer 2
 - Apply forward propagation to make final predictions.
 1. input \rightarrow x, opt θ_{1} , opt θ_{2}
- 10) Alpha is specified (0.75) \rightarrow learning rate
 11) Classifier is called to train neural network
 1 It takes input: array 'X', classification of outputs, alpha, IEPILON.
 2 It specifies some variables

- THETA1 = 0
- THETA2 = 0
- THETA3 = 0
- FC = 0
- J = 0

3 A function “nn” is called.

- It takes (X, Y, THETA1, THETA2, THETA3, IEPSILON, initw(initial state), alpha, J, FC) as input
- Thetas are calculated through following formula: -
- $THETA1 = 2 * IEPSILON * \text{rand}(K, \text{number of features}) - IEPSILON$
- $THETA2 = 2 * IEPSILON * \text{rand}(K, K+1) - IEPSILON$
- $THETA3 = 2 * IEPSILON * \text{rand}(K, K+1) - IEPSILON$
- In a loop forward propagation is applied and if it is the ‘500’ iteration then some calculations are done.
- 1. This forward propagation takes input X, THETA1 THETA2, THETA3.
- Then in the same loop back propagation is done through function “backpropclassification”
 1. Input $\rightarrow Y, A3, A2, \alpha, N$
THETA3, THETA2.
 2. Output $\rightarrow DELTA1, DELTA2, DELTA3$.
- Then in the same loop new theta values are calculated: -
 $THETA1_NEW = THETA1 - (\alpha * DELTA1)$
 $THETA2_NEW = THETA2 - (\alpha * DELTA2)$
 $THETA3_NEW = THETA3 - (\alpha * DELTA3)$

- 12) Test month \rightarrow Testing results to make future predictions about weather features
- 13) Save yearly predicted data into text files.
- 14) Plot graphs for weather features of 20 years.

IV. RESULTS AND DISCUSSION

After designing and training the multi-layered neural network with the existing dataset, we have obtained a relationship between the existing non-linear parameters of weather. The result reveals that by increasing the number of hidden layers, the trained neural network can classify and predict the weather variables with less error. This could even be possible to predict aforementioned weather parameters on

monthly bases, and daily bases. Due to the increasing number of figures in a single study, we have focused only on the classification of weather parameters and yearly based prediction of weather parameters. We have obtained the analytical results through above described methodology and these results are explained with detailed description along with each diagram. Each diagram illustrates the actual and predicted results for one year. Due to the increasing number of hidden layers, trained neural network can classify and predict the weather variables with less error.

Fig. 3 graph, shows actual weather (sunny, foggy, thunderstorm, and rainy) and weather specified after classification by neural network over the course of 20 years (7000 days). Fig. 4 graph, shows actual and predicted mean wind speed over the course of 20 years (7000 days).

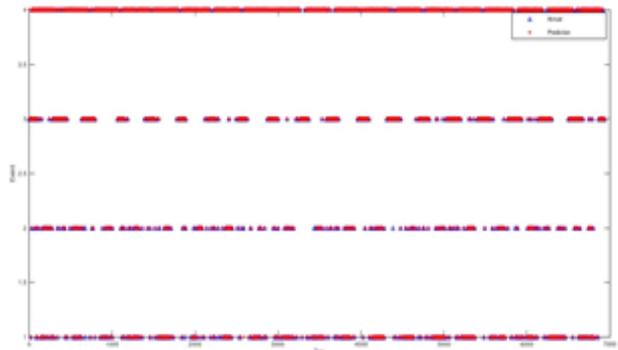


Fig. 3. Weather Event Classification.

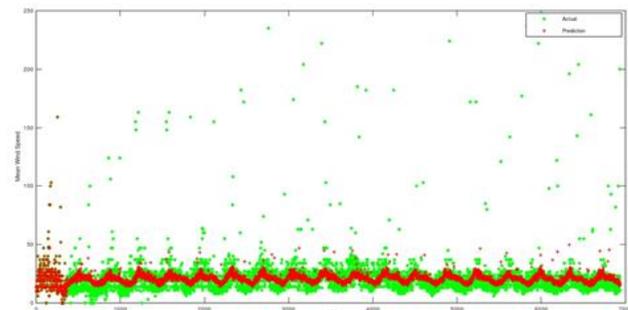


Fig. 4. Actual and Predicted Minimum Wind Speed.

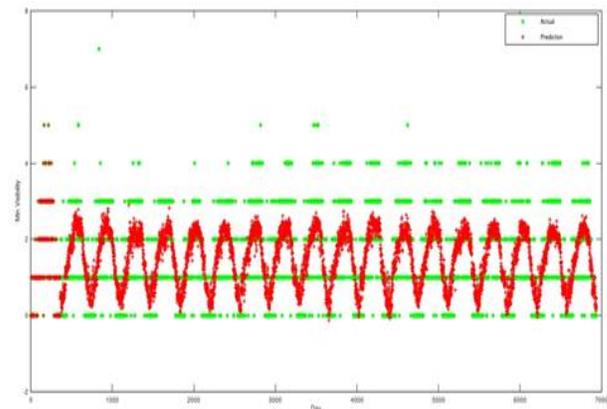


Fig. 5. Actual and Predicted Minimum Visibility.

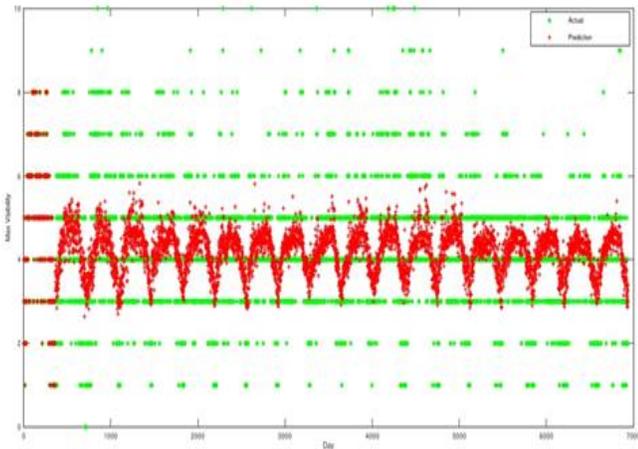


Fig. 6. Actual and Predicted Maximum Visibility.

Fig. 5 graph shows actual and predicted min visibility over the course of 20 years (7000 days). Fig. 6 graph that shows actual and predicted max visibility over the course of 20 years (7000 days).

Fig. 7 graph that shows actual and predicted min pressure over the course of 20 years (7000 days). Fig. 8 graph that shows actual and predicted max pressure over the course of 20 years (7000 days).

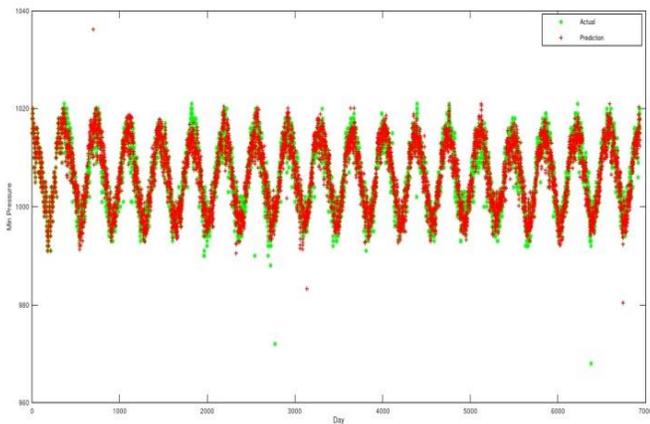


Fig. 7. Actual and Predicted Minimum Pressure.

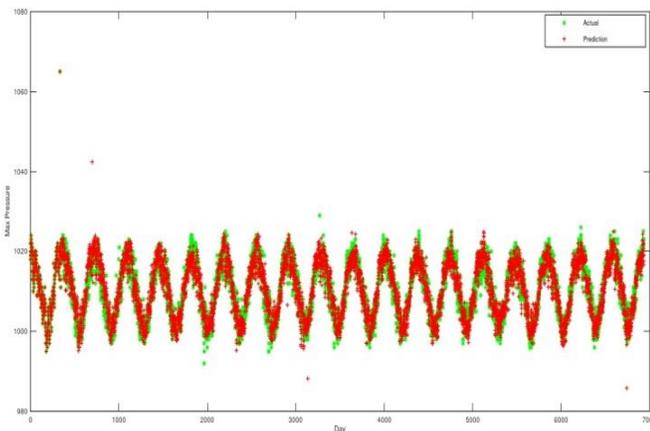


Fig. 8. Actual and Predicted Maximum Pressure.

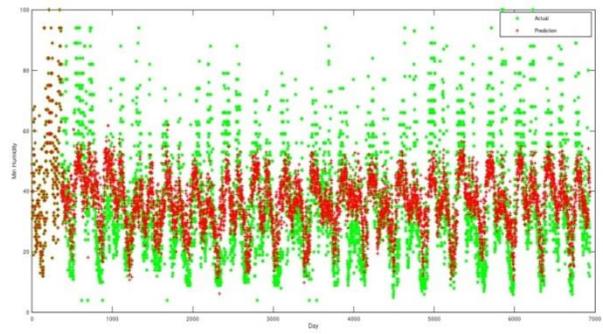


Fig. 9. Actual and Predicted Minimum Humidity.

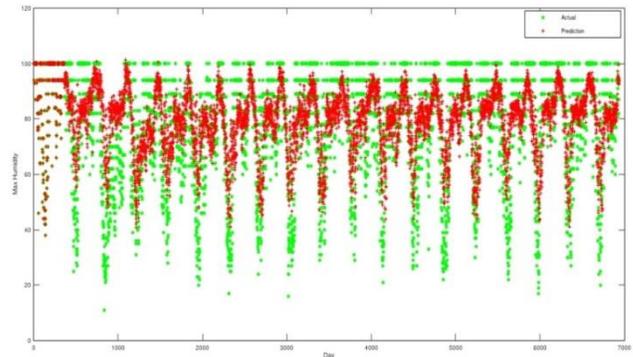


Fig. 10. Actual and Predicted Maximum Humidity.

Fig. 9 graph shows actual and predicted min humidity over the course of 20 years (7000 days). Fig. 10 graph shows actual and predicted max humidity over the course of 20 years (7000 days).

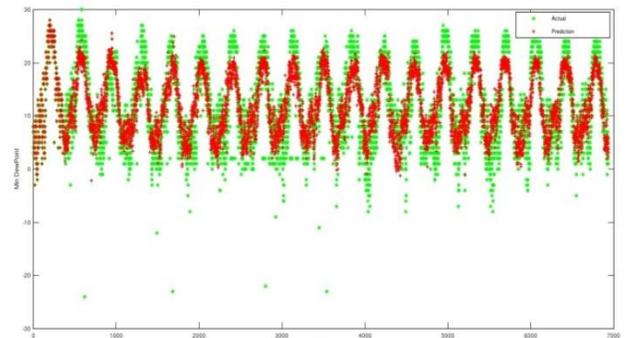


Fig. 11. Actual and Predicted Minimum Dewpoint.

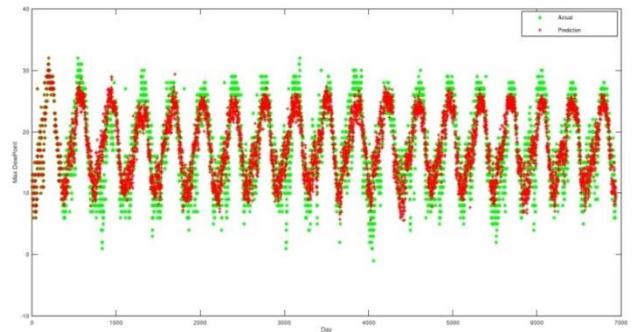


Fig. 12. Actual and Predicted Maximum Dewpoint.

Fig. 11 graph that shows actual and predicted min dewpoint over the course of 20 years (7000 days). Fig. 12 graph that shows actual and predicted max dewpoint over the course of 20 years (7000 days).

Fig. 13 graph that shows actual and predicted min temperature over the course of 20 years (7000 days). Fig. 14 graph that shows actual and predicted max temperature over the course of 20 years (7000 days).

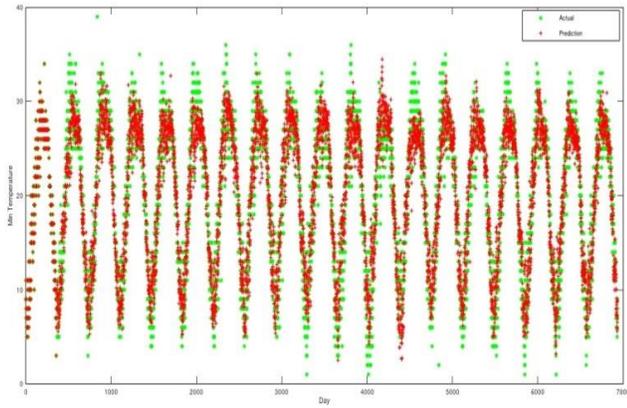


Fig. 13. Actual and Predicted Minimum Temperature.

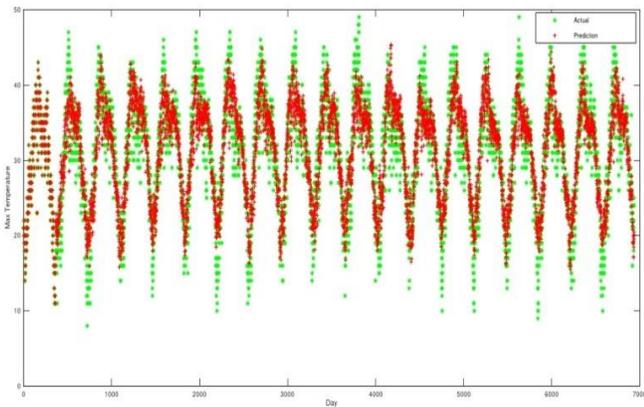


Fig. 14. Actual and Predicted Maximum Temperature.

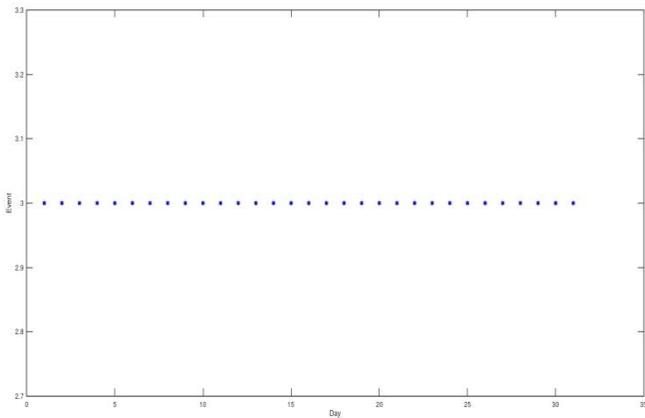


Fig. 15. Weather event Classification for One Month.

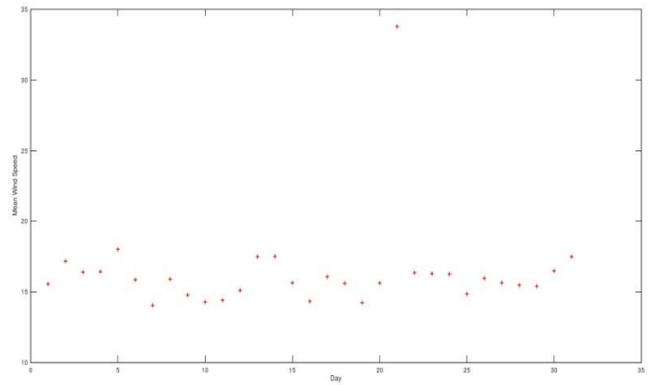


Fig. 16. Mean Wind Speed in Month Jan 2016.

Fig. 15 graph of classification of weather into rainy, thunderstorm, foggy, sunny (here these categories are converted to numbers and then categorized. The classification is done through the neural network which is trained by 20 years of data) in year 2016 over the course of 365 days. Fig. 16 graph of prediction of mean wind speed in year 2016 over the course of 365 days.

Fig. 17 graph of prediction of min visibility in year 2016 over the course of 365 days. Fig. 18 graph of prediction of max visibility in year 2016 over the course of 365 days.

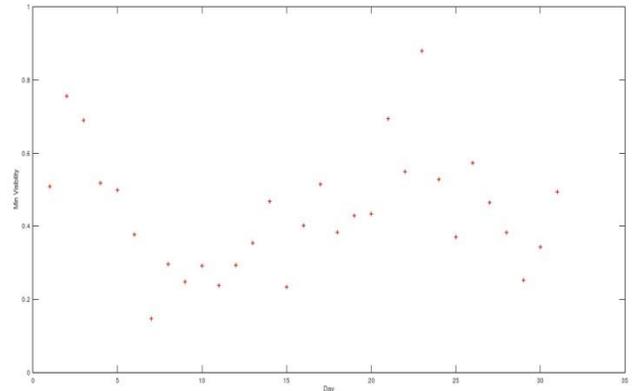


Fig. 17. Min Visibility in Month Jan 2016.

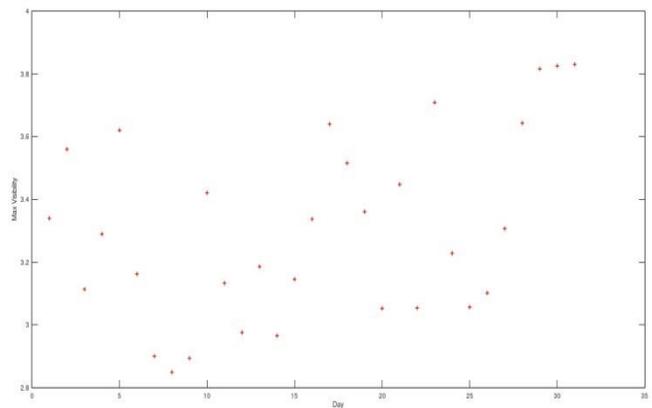


Fig. 18. Max Visibility in Month Jan 2016.

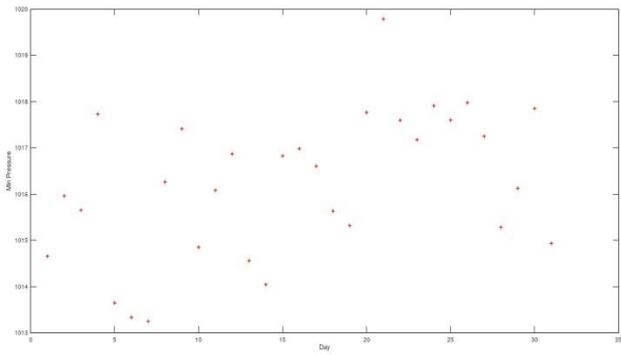


Fig. 19. Min Pressure in Month Jan 2016.

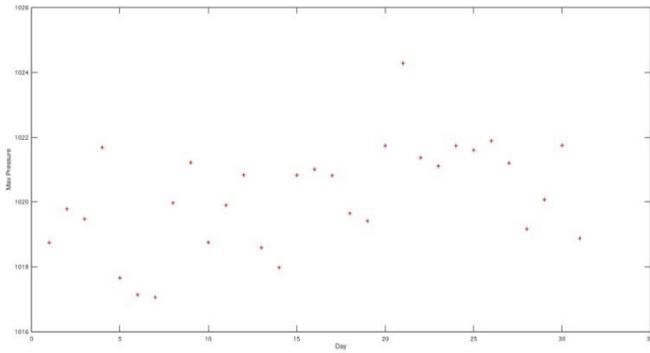


Fig. 20. Max Pressure in Month Jan 2016.

Fig. 19 graph of prediction of min pressure in year 2016 over the course of 365 days. Fig. 20 graph of prediction of max pressure in year 2016 over the course of 365 days.

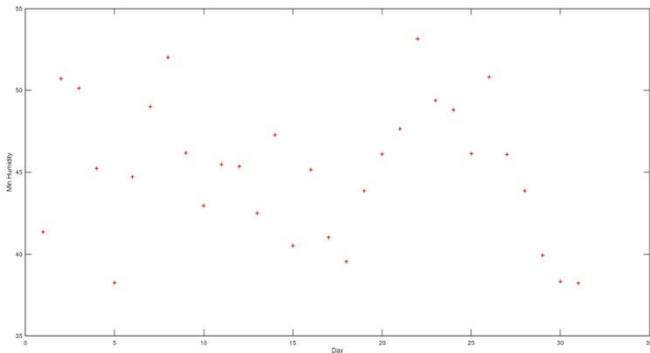


Fig. 21. Min Humidity in Month Jan 2016.

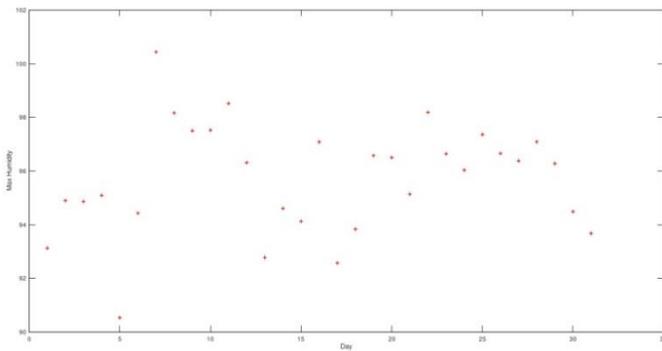


Fig. 22. Max Humidity in Month Jan 2016.

Fig. 21 graph of prediction of min humidity in year 2016 over the course of 365 days. Fig. 22 graph of prediction of max humidity in year 2016 over the course of 365 days.

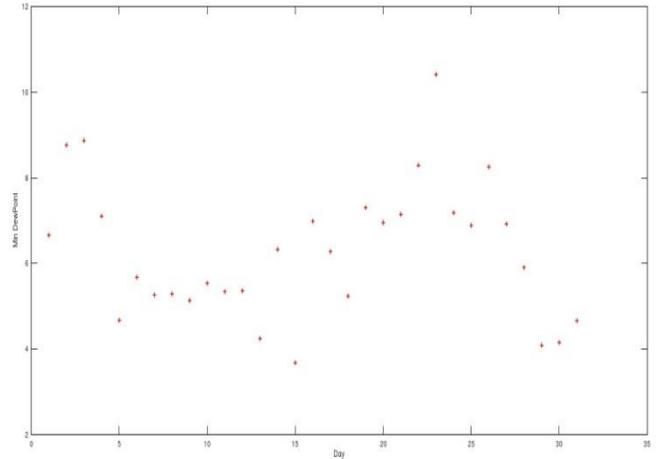


Fig. 23. Min DewPoint in Month Jan 2016.

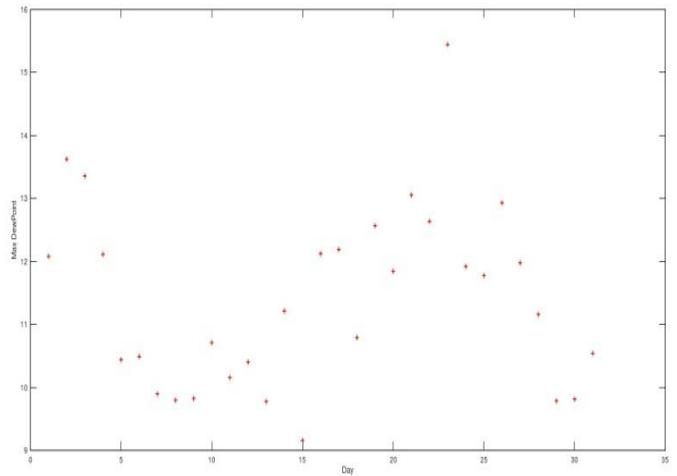


Fig. 24. Max DwePoint in Month Jan 2016.

Fig. 23 graph of prediction of min dewpoint in year 2016 over the course of 365 days. Fig. 24 graph of prediction of max Dewpoint in year 2016 over the course of 365 days.

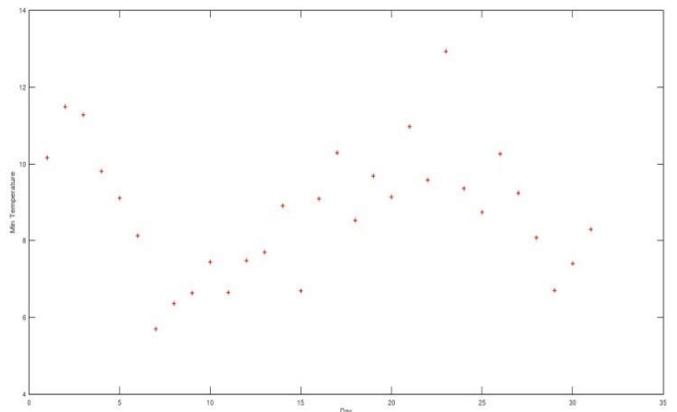


Fig. 25. Min Temperature in Month Jan 2016.

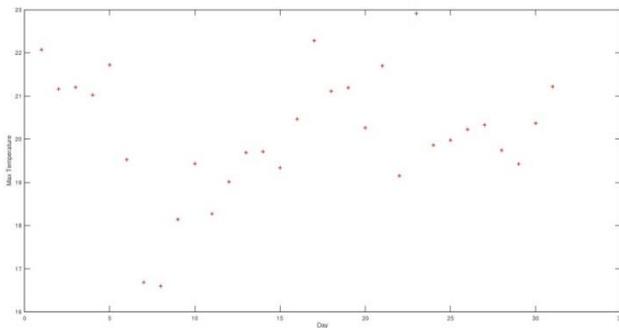


Fig. 26. Max Temperature in Month Jan 2016.

Fig. 25 graph of prediction of min temperature in year 2016 over the course of 365 days. Fig. 26 graph of prediction of max temperature in year 2016 over the course of 365 days.

V. CONCLUSION

In this paper, a most powerful prediction algorithm called back propagation algorithm was used to predict and classify the weather forecast standard dataset. Eleven weather features were used to perform classification of weather into four types. Furthermore, twenty training examples from 1997-2015 were used to predict eleven weather features. The prediction was calculated for weather forecast basic factors like humidity, speed, etc. A Multi-layered neural network is designed and trained with the existing dataset and obtained a relationship between the existing non-linear parameters of weather. The overall behavior of our model has been concluded is that by increasing the number of hidden layers, the trained neural network can classify and predict the weather variables with less error. In upcoming paper, we will try to predict aforementioned weather parameters on monthly bases, and daily bases. Due to increasing number of figures in a single study, we have focused only on two objectives, i.e, classification and year based prediction. Additionally, a comparative analysis of several ANN techniques will be made so that we can emphasis our results broadly.

REFERENCES

- [1] Delle Monache, Luca, Thomas Nipen, Yubao Liu, Gregory Roux, and Roland Stull. "Kalman filter and analog schemes to postprocess numerical weather predictions." *Monthly Weather Review* 139, no. 11 (2011): 3554-3570.
- [2] Delle Monache, Luca, F. Anthony Eckel, Daran L. Rife, Badrinath Nagarajan, and Keith Searight. "Probabilistic weather prediction with an analog ensemble." *Monthly Weather Review* 141, no. 10 (2013): 3498-3516.
- [3] De Giorgi, Maria Grazia, Antonio Ficarella, and Marco Tarantino. "Assessment of the benefits of numerical weather predictions in wind

- power forecasting based on statistical methods." *Energy* 36, no. 7 (2011): 3968-3978.
- [4] Malik, Pooja, Saranjeet Singh, and Binni Arora. "An effective weather forecasting using neural network." *Int J Emerg Eng Res Technol* 2, no. 2 (2014): 209-212.
- [5] Olaiya, Folunsho, and Adesesan Barnabas Adeyemo. "Application of data mining techniques in weather prediction and climate change studies." *International Journal of Information Engineering and Electronic Business* 4, no. 1 (2012): 51.
- [6] de Lima, Glauston R. Teixeira, and Stephan Stephany. "A new classification approach for detecting severe weather patterns." *Computers & geosciences* 57 (2013): 158-165.
- [7] Sun, Xiaorong, Peter B. Luh, Kwok W. Cheung, Wei Guan, Laurent D. Michel, S. S. Venkata, and Melanie T. Miller. "An efficient approach to short-term load forecasting at the distribution level." *IEEE Transactions on Power Systems* 31, no. 4 (2016): 2526-2537.
- [8] Anandharajan, T. R. V., G. Abhishek Hariharan, K. K. Vignajeth, and R. Jijendiran. "Weather Monitoring Using Artificial Intelligence." In *Computational Intelligence and Networks (CINE), 2016 2nd International Conference on*, pp. 106-111. IEEE, 2016.
- [9] Araújo, Ricardo de A., Adriano LI Oliveira, Sergio Soares, and Silvio Meira. "Dilation-erosion perceptrons with evolutionary learning for weather forecasting." In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pp. 3070-3077. IEEE, 2011.
- [10] Islam, Taohidul, Sajal Saha, Ali Ahmed Evan, Nabonita Halder, and Shakti Chandra Dey. "Monthly Weather Forecasting through ANN Model: A Case Study in Barisal, Bangladesh." *International Journal of Advanced Research in Computer and Communication Engineering* 5, no. 6 (2016).
- [11] Chen, Changsong, Shanxu Duan, Tao Cai, and Bangyin Liu. "Online 24-h solar power forecasting based on weather type classification using artificial neural network." *Solar Energy* 85, no. 11 (2011): 2856-2870.
- [12] Purnomo, H. D., K. D. Hartomo, and S. Y. J. Prasetyo. "Artificial neural network for monthly rainfall rate prediction." In *IOP Conference Series: Materials Science and Engineering*, vol. 180, no. 1, p. 012057. IOP Publishing, 2017.
- [13] Detwiler, Luke. "Using Backpropagation Neural Networks for the Prediction of Residual Shear Strength of Cohesive Soils." (2014).
- [14] Narvekar, Meera, and Priyanca Fargose. "Daily weather forecasting using artificial neural network." *International Journal of Computer Applications* 121, no. 22 (2015).
- [15] Balsamo, Gianpaolo, Rui Salgado, Emanuel Dutra, S. Boussetta, T. Stockdale, and Miguel Potes. "On the contribution of lakes in predicting near-surface temperature in a global weather forecasting model." *Tellus A: Dynamic Meteorology and Oceanography* 64, no. 1 (2012): 15829.
- [16] Viswambari, M., and Dr R. Anbu Selvi. "Data Mining Techniques to Predict Weather: A Survey." *International Journal of Innovative Science, Engineering & Technology* 1, no. 4 (2014): 3.
- [17] Sukanya, R., and K. Prabha. "Comparative Analysis for Prediction of Rainfall using Data Mining Techniques with Artificial Neural Network." Volume-5, Issue-6, Page 288-292 (2017).
- [18] Maleki, Akbar, Morteza Gholipour Khajeh, and Marc A. Rosen. "Weather forecasting for optimization of a hybrid solar-wind-powered reverse osmosis water desalination system using a novel optimizer approach." *Energy* 114 (2016): 1120-1134.

An Incremental Technique of Improving Translation

Aasim Ali

Department of Computer Science
Bahria University
Lahore, Pakistan

Arshad Hussain

Department of Electrical Engineering
University of Central Punjab
Lahore, Pakistan

Abstract—Statistical machine translation (SMT) refers to using probabilistic methods of learning translation process primarily from the parallel text. In SMT, the linguistic information such as morphology and syntax can be added to the parallel text for improved results. However, adding such linguistic matter is costly, in terms of time and expert effort. Here, we introduce a technique that can learn better shapes (morphological process) and more appropriate positioning (syntactic realization) of target words, without linguistic annotations. Our method improves result iteratively over multiple passes of translation. Our experiments showed better accuracy of translation, using a well-known scoring tool. There is no language specific step in this technique.

Keywords—Statistical machine translation; incremental learning algorithm; English; Urdu

I. INTRODUCTION

Recent trend in machine translation is mostly towards data-driven methods including Statistical Machine Translation (SMT), which uses parallel text. This approach learns translation through phrase alignments [1] which are based on word alignments. In SMT, the morphological information improves learnability for realizing the correct shape of words, especially for morphologically rich languages like Arabic and Urdu. Similarly, the syntactic information improves positioning of words in the given context, especially when source and target pair has different positions for grammatical relations (Subject, Object, etc.) like English versus Urdu. An intuitive way of algorithmic evaluation of translation output is based on the number of matching sequences and subsequences of words in comparison with human translation. We have used BLEU [2] for an automatic evaluation of progress in translation improvement. A freely available toolkit for training and decoding of SMT systems, Moses [3] is used in our experiments, along with the supportive tools [4] for intermediate tasks like text alignment. Open source tools [5] are used for English (the source side of parallel text), and locally developed morphology analyzer [6] and POS tagger [7] of Urdu (target side of parallel text) are used for morpho-syntactic experiment. The experiments for baseline and proposed technique, both, use plain parallel text.

In the proposed method, the system gradually learns these linguistic elements (shapes and orders of words, etc.) from the surface forms of the target side, without any explicit knowledge, hint and tagging. There is no need of mono-lingual resources either, in addition to the parallel text. We have

improved the shapes and arrangements of words on the target side by using the SMT process iteratively, to incrementally learn such information from simply the parallel text itself.

The rest of this paper has been organized in the following sections. Section II gives a review of the existing work on the statistical machine translation and the incremental learning. Section III details the methodology of the proposed algorithm. Section IV describes the data, experimental setup, and results. Section V discusses the proposed technique in the light of obtained results.

II. LITERATURE REVIEW

Statistical machine translation [8], being a machine learning approach towards translation [9], is used in the proposed work. A more detailed and updated record of statistical machine translation may be found in [8]. The proposed work considers linguistic knowledge (morphology, syntax, and word sense) to be “hidden” elements and uses the iterations of machine translation in the form of expectation maximization algorithm [10] without any external knowledge, to reach a better output. Words in our output are better in terms of correctness of shapes, sequences, and senses. The proposed work considers the intermediate translation of source as a pivot language [11], which is then used to improve the model to gradually reach the target language, by utilizing the power of incremental learning [12; 13; 14]. Gradual learning in several iterations reduces the impact of noise and irrelevant attributes [15] for automatically learning the word mappings to generate more correct sentences as output of translation.

The approach of incremental machine translation [16] uses the knowledge of human translator for enhancing the confidence of correct translations, and using that confidence for future translations. The proposed work uses the same idea of enhanced confidence with the help of an automatic tool, BLEU, for evaluation of translated output of one pass to be used as input for translation of next pass. Daybelge and Cicekli [17] have used a similar approach of using BLEU score as a measure of incremental learning and reported improvement in the translation quality using example based machine translation. Quality of translation does not depend only on the syntax and morphology but also on the sense of the source word [18; 19]. Using the translation of phrases observed previously increases the translation correctness when they occur subsequently [20; 21]. This is another view of “incremental” learning in which already observed high probability mappings help improving the mappings of other translation units in subsequent passes of learning. We have successfully experimented and introduced a technique that

This paper is out of my PhD work at NUCES (Lahore Campus) which is partially funded by HEC (Pakistan).

gradually learns the linguistic information from parallel text in several iterations of translation, which is detailed in the next section.

```
1  $x \leftarrow 0$ 
2  $Diff \leftarrow 0$ 
3  $i \leftarrow 1$ 
4  $Bscore_0 \leftarrow 0$  // considering that two texts
5 // (source and target) are disjoint
6 do
7 {
8    $Model_i \leftarrow SMT\_Learner(TT_{i-1}, TT_n)$ 
9    $THOT_i \leftarrow SMT\_Decoder(Model_i, THOT_{i-1})$ 
10   $Bscore_i \leftarrow BLEU\_Score(THOT_i, THOT_n)$ 
11   $Diff \leftarrow Bscore_i - Bscore_{i-1}$ 
12   $TT_i \leftarrow SMT\_Decoder(Model_i, TT_{i-1})$ 
13 } while ( $Diff > x$ )
```

Fig. 1. Algorithm for Incremental Learning.

III. ALGORITHM FOR INCREMENTAL LEARNING

Fig. 1 shows the complete algorithm of incremental learning. The labels and variables used in the following algorithm are defined as: $Bscore_i$ means BLEU Score of i th iteration; $Diff$ means the difference of two consecutive $Bscore$ values to be compared with the Threshold (that is x); $Model_i$ is the SMT model learnt in i th iteration; when $i=1$ then TT_{i-1} (TT_0) means Training Text which is source side, and TT_i (TT_1) means translation of source side, same goes for all values of i ; TT_n means the target side for learning next SMT model; $THOT_0$ is the source side of held-out text, $THOT_i$ denotes the i th translated version of the source side of held-out text; and $THOT_n$ denotes target side of held-out text.

Line 1 and **2** initialize the variables x and $Diff$ to 0. **Line 3** initializes the iteration counter i to 1. **Line 4** initializes the BLEU score variable $Bscore_0$ to 0, which means $Bscore$ for 0th iteration is 0. This variable will be used to compute the improvement in the translation for comparison with the $Bscore$ of i th iteration for measuring the threshold. **Line 6** to **13** is a loop that will continue for the specific threshold. In this instance of the algorithm, the loop will stop when there is no improvement in the $Bscore$, because the threshold testing variable x is kept 0.

Inside this loop, **line 8** updates the SMT Model for i th iteration, between the TT_{i-1} and TT_n . When $i = 1$, for first iteration, then TT_0 is the original training text on the source side of translation (English in our case, see section 4). When $i > 1$, for subsequent iterations, then TT_1 , TT_2 , and so on, are the i th translated versions of source training text; thus termed as translated text. TT_n is always the original training text on the target side of translation (Urdu in our case, see section 4). Hence, in the first iteration we obtain $Model_1$ which is trained SMT model for translation from English into Urdu. In **line 9**, the original held-out text on the source side ($THOT_0$) is translated using $Model_1$ thus generating the translated version of held-out text ($THOT_1$), when $i = 1$. When $i > 1$ then every $THOT_i$ is the translated version of $THOT_{i-1}$ using $Model_i$. The

line 10 computes the BLEU score between translated version and the target side of the held-out data. **Line 11** subtracts the BLEU score of previous iteration ($Bscore_{i-1}$) from the BLEU score of current iteration ($Bscore_i$). When $i = 1$ then $Bscore_0$ is 0 and $Bscore_1$ is the BLEU score of first iteration. Thus, TH holds the difference between $Bscore_i$ and $Bscore_{i-1}$ for every iteration. The processing of **line 12** produces the translation of source side of training text which may have to be used in the subsequent iteration if the loop continues to next iteration. TT_i is the translated version of TT_{i-1} using $Model_i$. When $i = 1$ then TT_1 is the translation of original source text TT_0 . When $i > 1$ then every TT_i is the translation of corresponding TT_{i-1} . If there is no gain in the $Bscore$ then the value of TH remains equal to or less than 0 thus the loop terminates. As the final output of this algorithm we obtain SMT Models from $Model_1$ to $Model_n$, from parallel training data. Our stopping criteria depends on the held-out data; and we use all these models in an incremental way to decode the evaluation data (test data). All these datasets are distinct for our experiment.

IV. VERIFYING EXPERIMENT

The baseline experiment is performed by learning simple phrase based machine translation (PBMT) [22] from plain parallel text. Next, we added POS tags and morphological annotations as factors in the factor based [23] PBMT, to see the improved result. Then we trained using our proposed model to achieve the best result. The data is described below in subsection A, and the experiments are detailed in subsection B, of this section.

A. Data

Text from two books is used in this study. The English and Urdu versions of these books are already aligned at topic level (containing one or more paragraphs). There are 497,354 words in 26,822 sentences on English side and 513,550 words in parallel Urdu translations.

We partitioned our data into three disjoint segments: 75% as training data, 19% as held-out data, and 6% as evaluation data. Plain bi-text is used for baseline and for incremental learning. The lemma, morphological tags and POS on source side (English) are computed using open source tools [5]. Similar tools for the target side (Urdu) of the parallel corpus are developed locally. The finite state transducer [6; 24] is used for morphology, and TNT tagger [7; 25] is used for POS tagging.

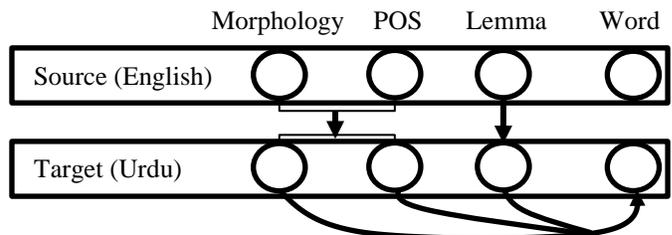


Fig. 2. Mapping of Factors.

Factored translation model is used for incorporating linguistic information at word level. Each such additional information attached to a word is termed as factor. These factors are used in a series of mapping stages or steps. The steps may be of two types: (a) translating the factors on input side to those on the output side, and (b) using the existing factors on output side to generate other factors on the same side for rendering the final shape of the word. Fig. 2 shows the mapping of factors.

The following mapping of input/output factors is used:

- 1) Lemma of input side is translated into the lemma on the output side.
- 2) POS and morphological factors on the input side are also translated into the factors on the output side.
- 3) Surface form on the output side is generated using the translated morpho-syntactic factors on the output side.

B. Experiment and Result

First of all, plain bi-text is used to obtain the baseline results. Then the same model is tuned for held-out data using minimum error rate training [26], which improved results from 32.10 to 37.10. Then the morpho-syntactic model of translation is used for which words are annotated with lemma and POS tag factors. This experiment produced the BLEU score of 36.73.

The proposed technique of incremental learning is designed to test if the un-annotated text can itself incrementally take the desired shapes and sequences of words induced by the implicit morpho-syntactic knowledge which is always present in the running text. The proposed algorithm is implemented in the following way:

- 1) Executed the training model of baseline, i.e. Source-to-Target Model ($Model_1$), on the training set (TT_0) itself (to prepare an intermediate train set TT_1). The translation of held-out data ($THOT_0$) from $Model_1$ is also saved and termed as $THOT_1$ to be used in the next stage.
- 2) Used that translated training part (TT_1) to pair with the target side (TT_n) of the corpora to learn a new model ($Model_2$) to automatically learn the good mappings which were missed in the first pass (while learning the $Model_1$).
- 3) Used $Model_2$ on $THOT_1$ to obtain the next version of translated held-out data ($THOT_2$), and found the improvement in the BLEU score by comparing between $THOT_2$ and $THOT_n$.
- 4) Executed $Model_2$ on the TT_1 (to prepare another intermediate train set TT_2) for next stage of learning.
- 5) Then used that latest translated training part (TT_2) to pair with the target side (TT_n) of the corpora to learn another model ($Model_3$) to further learn the good mappings which were missed even in the second pass.
- 6) Then executed $Model_3$ on $THOT_2$ to obtain $THOT_3$ and found the improvement in the BLEU score by comparing between $THOT_3$ and $THOT_n$.
- 7) Finally, for the sake of evaluation on a data set which is kept separately (apart from training and held-out data sets), executed $Model_1$ on original source side of the evaluation data

(ET_0) to obtain ET_1 . Then executed $Model_2$ on ET_1 to obtain ET_2 . Afterwards executed $Model_3$ on ET_2 to obtain ET_3 , which produced the highest BLEU score from ET_3 versus ET_n . The detail of this step is shown in Fig. 3.

In Fig. 3, each rectangle represents the process, each parallelogram signifies an input/output of the process, each solid arrow shows the sequence of flow, and each dashed arrow denotes the SMT model used in the process.

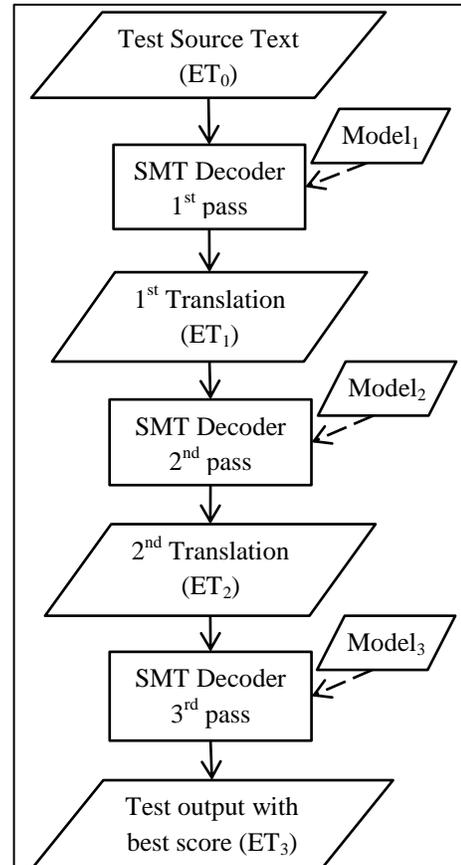


Fig. 3. Application of Three Models Learnt with Incremental Technique.

V. DISCUSSION AND CONCLUSION

The summary of results shown in Table I clearly shows that incremental learning proposed in this paper gives the highest BLEU score. One reason of unprecedentedly high score under proposed technique is the significant overlap of phrases in the data. However, it is also important to keep in mind that gain from this overlap could not be exploited without using the power [12; 13; 14; 15] of incremental learning.

Since this approach involves no language-specific steps therefore it may be applied to any language pair. The technique of exploiting the overlapping in the training set, the held-out set and the evaluation set, may work well for translation of any other text that typically has significant overlap of phrases including user manuals, blogs, specific news genre, and research articles from a specific field. It may also be applied for word sense disambiguation [27] using parallel corpus, instead of using explicit linguistic knowledge to resolve the word sense.

TABLE I. RESULTS OF TRANSLATION OF EVALUATION SET

Experiment	BLEU
Translation with Trained Model ₁	32.10
Translation with Tuned Model ₁	37.10
Translation with Morpho-Syntactic Model	30.73
Translation with Model _{1,3} obtained from <i>Incremental Technique</i>	42.91

REFERENCES

- [1] Koehn, P., Och, F. J., & Marcu, D. (2003, May). Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 48-54). Association for Computational Linguistics.
- [2] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.
- [3] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 177-180). Association for Computational Linguistics.
- [4] Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19-51.
- [5] Carreras, X., Chao, I., Padró, L., & Padró, M. (2004, May). FreeLing: An Open-Source Suite of Language Analyzers. In LREC (pp. 239-242).
- [6] Ali, Aasim. (2010). Study of Morphology of Urdu Language, for Its Computational Modeling: Study of Morphological Patterns in Urdu Language, and Partial Implementation of Computational Solution for the Same Using a Finite State Tool. VDM Publishing.
- [7] Asif, T., Ali, A. and Malik, M. K. (2015). Developing a POS Tagged Resource of Urdu. *Science International*, 27(5), 4479-4483.
- [8] Koehn, P. (2010). *Statistical Machine Translation* (1st ed.). Cambridge University Press, New York, NY, USA.
- [9] Cardie, C., & Mooney, R. J. (1999). Guest editors' introduction: Machine learning and natural language. *Machine Learning*, 34(1), 5-9.
- [10] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.
- [11] Wu, H., & Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3), 165-181.
- [12] Kirby, S., & Hurford, J. (1997). The evolution of incremental learning: language, development and critical periods. *Edinburgh Occasional Papers in Linguistics*, 97(2), 1-33.
- [13] Giraud-Carrier, C. (2000). A note on the utility of incremental learning. *AI Communications*, 13(4), 215-223.
- [14] Solomonoff, R. J. (2002, December). Progress in incremental machine learning. In NIPS Workshop on Universal Learning Algorithms and Optimal Search, Whistler, BC.
- [15] Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36(2), 267-287.
- [16] Toselli, A. H., Vidal, E., & Casacuberta, F. (2011). Incremental and Adaptive Learning for Interactive Machine Translation. In *Multimodal Interactive Pattern Recognition and Applications* (pp. 169-177). Springer London.
- [17] Daybelge, T., & Cicekli, I. (2011). A ranking method for example based machine translation results by learning from user feedback. *Applied Intelligence*, 35(2), 296-321.
- [18] Lee, H. A. (2006). Translation selection through machine learning with language resources. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead* (pp. 370-377). Springer Berlin Heidelberg.
- [19] Carpuat, M., & Wu, D. (2007, June). Improving Statistical Machine Translation Using Word Sense Disambiguation. In *EMNLP-CoNLL* (Vol. 7, pp. 61-72).
- [20] Bannard, C., & Callison-Burch, C. (2005, June). Paraphrasing with bilingual parallel corpora. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 597-604). Association for Computational Linguistics.
- [21] Callison-Burch, C., Koehn, P., & Osborne, M. (2006, June). Improved statistical machine translation using paraphrases. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 17-24). Association for Computational Linguistics.
- [22] Zens, R., Och, F. J., & Ney, H. (2002, September). Phrase-based statistical machine translation. In *Annual Conference on Artificial Intelligence* (pp. 18-32). Springer, Berlin, Heidelberg.
- [23] Koehn, P., & Hoang, H. (2007). Factored translation models. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL).
- [24] Beesley, K. R., & Karttunen, L. (2003). *Finite-state morphology: Xerox tools and techniques*. CSLI, Stanford.
- [25] Brants, T. (2000, April). TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on Applied natural language processing (pp. 224-231). Association for Computational Linguistics.
- [26] Och, F. J. (2003, July). Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 160-167). Association for Computational Linguistics.
- [27] Specia, L., Srinivasan, A., Joshi, S., Ramakrishnan, G., & Nunes, M. D. G. V. (2009). An investigation into feature construction to assist word sense disambiguation. *Machine Learning*, 76(1), 109-136.

Role Term-Based Semantic Similarity Technique for Idea Plagiarism Detection

Ahmed Hamza Osman, Hani Moetque Aljahdali

Department of Information System

Faculty of Computing and Information Technology at Rabigh, King Abdulaziz University
Jeddah, Saudi Arabia

Abstract—Most of the text mining systems are based on statistical analysis of term frequency. The statistical analysis of term (phrase or word) frequency captures the importance of the term within a document, but the techniques that had been proposed by now still need to be improved in terms of their ability to detect the plagiarized parts, especially for capturing the importance of the term within a sentence. Two terms can have a same frequency in their documents, but one term pays more to the meaning of its sentences than the other term. In this paper, we want to discriminate between the important term and unimportant term in the meaning of the sentences in order to adopt for idea plagiarism detection. This paper introduces an idea plagiarism detection based on semantic meaning frequency of important terms in the sentences. The suggested method analyses and compares text based on a semantic allocation for each term inside the sentence. SRL offers significant advantages when generating arguments for each sentence semantically. Promising experimental has been applied on the CS11 dataset and results revealed that the proposed technique's performance surpasses its recent peer methods of plagiarism detection in terms of Recall, Precision and F-measure.

Keywords—Plagiarism detection; semantic similarity; semantic role; term frequency; idea

I. INTRODUCTION

Given the bigness of the online, plagiarism, or the intended use of somebody else's original data while not acknowledge its supply, has been a heavy drawback in areas like Literature, Science, and Education. The convenience of access to proprietary contents has become an issue of concern additionally for scholars. The challenge is exacerbated when the suspected text generated semantically, which is known as idea plagiarism. It is not solely the extra problem of manually capturing the concept or idea performed, however additionally the people's lack of information concerning writing ethical issues and text paraphrasing. The different categories of plagiarism are cut-and-paste, ideas plagiarism, semantic plagiarism and paraphrasing, style plagiarism, authorship and citation plagiarism [1]. Several works had been done in text plagiarism detection based on the lexical and syntactic structure of the writing and failed to detect the semantic and idea plagiarism. However, most of these methods are created for verbatim duplicates, and similarity performance is decreased when dealing with plagiarism with heavy cases [2], due to paraphrasing and semantic similarity cases. Recently, different studies tried to develop and improve the accurate methods in semantic and idea text plagiarism detection domain

such as [3-6] Alzahrani et al. [7]; Maurer et al. [4]; Gupta and Deep [6]; 2016; Vani and Gupta [5]; Juan D. Velásquez and et al. [8]; Weber-Wulff [9]).

The rest of the paper is ordered as follows: related work and Literature review in different type of plagiarism detection is considered in Section 2. In Section 3, a suggested proposed solution is presented. Section 4 discussed a full depiction of the idea plagiarism detection and role-based similarity that formulated in the proposed method. Corpus and experimental design that conducted in the suggested method is described in Section 5. In Section 6, output results and discussion is provided, whereas Section 7 is devoted to the conclusion.

II. MATERIAL AND METHODS

Several studies have discussed plagiarism in academia filed [4, 10, 11], and demonstrated different categories of available plagiarism detection techniques. For instance, Vani and Gupta[5] proposed an idea plagiarism detection method based on semantic syntax concept extraction. The extracting of the concepts was generated using a genetic algorithm. Their method detects the idea plagiarism based on two level Document level and word level. They tried to combine the similarity measure that employs the semantic concept extraction and then used for passage stage matching [5].

Palkovskii, Belov and Muzika [12] presented Exploring Fingerprinting as an outside plagiarism identification strategy to PAN-PC-2010. Their framework was initially created as a component of the proposal stockpiling framework utilized by the Zhytomyr State University. Palkovskii, Belov, and Muzika depended on fingerprinting and hash look techniques for finding likenesses between reports.

Osman and et al. [1] proposed a detection scheme based on SRL and concept extraction. The SRL used for extracting the roles and arguments for each sentence and the wordNet used to extract the sense of each term inside the sentence. The proposed method can use in different type of detection such as copy and paste, semantic and paraphrasing, structure plagiarism [1].

Palkovskii, Belov, and Muzyka, [13] likewise returned contenders. Like Encoplot, they proposed a WordNet-based semantic similitude estimation for the outside counterfeiting identification shown in PAN-PC-09 and was enhanced for PAN-PC-2010. For PAN-PC-2011, they exhibited a gauge for further review by demonstrating unmistakably characterized corpus measurements, for example, outside and inherent,

confusion techniques, point coordinate, case length and report length. They demonstrated that an immediate connection could be made between the complication technique and accomplished execution.

Sheffield University spoke to by Parth, Sameer and Majumdar, [14] thought of a framework that was intended to identify extraneous counterfeiting. They utilized a three-organize technique for pre-preparing; record choice utilizing term n-grams and their last examination utilized a Running Karp-Rabin Greedy String Tiling string coordinating method. Their framework was granted a score of 0.20 for general execution with scores of 1.21 for Granularity, 0.16 for the Recall measure and 0.4 for Precision.

Chong Specia and Mitkov [15], proposed another system for unoriginality location given the string coordinating and Naïve-Bayes classifier. Guileless Naïve-Bayes algorithm is a straightforward classifier given the Bayes' hypothesis. The technique is accustomed to creating a probabilistic framework of the plagiarism short answer questions dataset. The point of a Naïve-Bayes algorithm is to take on their framework to order different cases given the dataset classes (cut-and-paste, light, heavy, and non-plagiarized). The strategy chose the arrangement of "best components" and utilized them to take in their framework to group a given potential plagiarism record as having a place with one of the plagiarism classes.

We show through the previous studies a various works has been done to detect and capture the text plagiarism. However, these approaches that have been introduced still need to be enhanced to finding the academic and scientific plagiarized idea, mainly in semantic structure part.

III. PROPOSED METHOD

It's very significant to note that exploiting relationship between terms roles and their verbs in the same text has promising possible for understanding the idea of the text. The ordinary method of text mining is to capture term frequency. In this paper, Semantic Role-based model for text similarity and idea plagiarism detection will be introduced. The main concept of the suggested technique is capturing the meaning construction of sentence terms within a text and documents; capturing and role terms frequency; and capturing the idea similarity based on semantic structure and role terms frequency together. The proposed method compromise of different phases starting from the preprocessing steps such as sentences chunking and stop term removal. Then, the semantic role labeling technique will be utilized to exploit the roles of all terms inside the sentences. The last steps are text similarity and idea extraction based on role frequency and role term similarity. The main framework of the suggested technique is shown in Fig. 1.

Every one of these means will be further talked about in the accompanying segments:

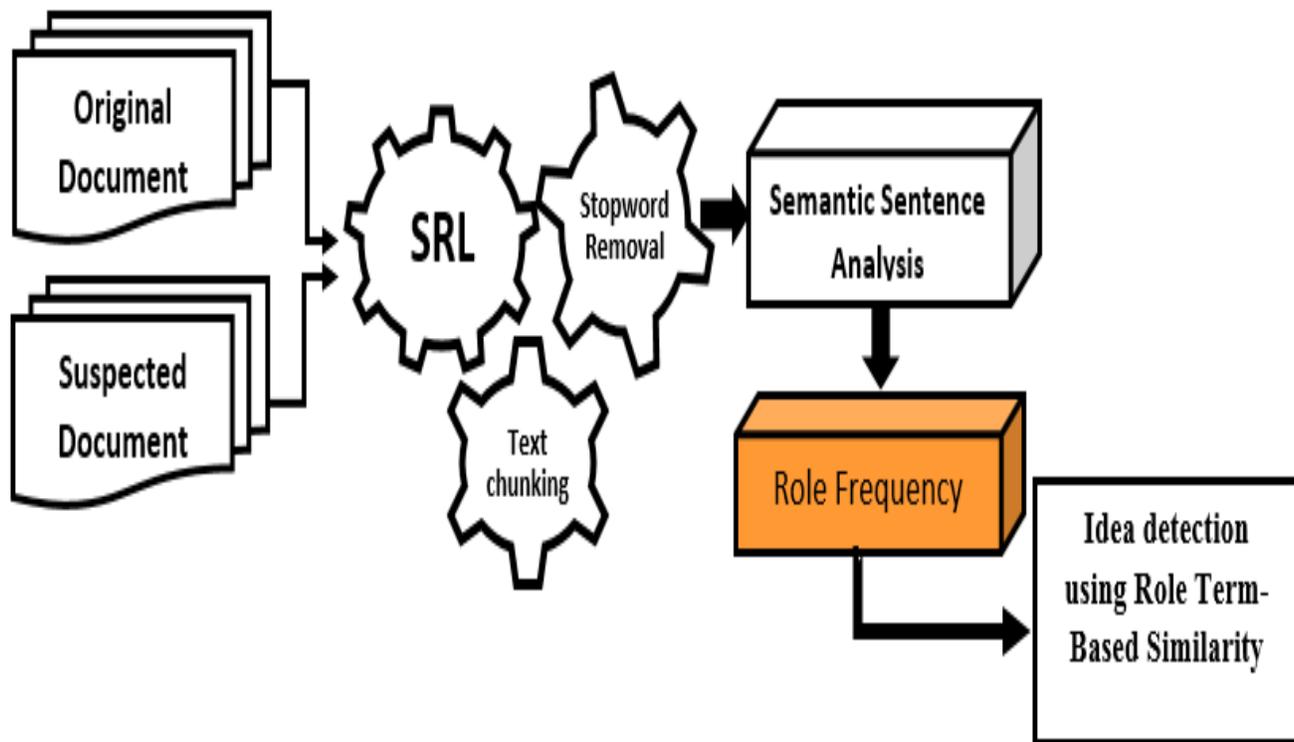


Fig. 1. Proposed Framework.

A. Semantic-Role Labeling (SRL)

Semantic constructions or linguistics frames were suggested by Fillmore [16] wherever general frames were utilized for general themes and roles such as and PropBank introduced by Palmer et al. [17]. FrameNet suggested by Baker et al. [18]. A statistical scheme is learned on the data from the FrameNet project to mechanically assign linguistics roles [17]. Pradhan et al.[19], Surdeanu et al. [20], and Xue and Palmer [21] followed this method by enhancing sets of data mining approaches. Barnickel et al. [22] proposed a large scale system based on SRL and neural network for extracting relation from biomedical data. This technique essentially used SENNA tool that is utilized in different text processing system. The SENNA tool can exploit the terms roles for each sentence based on the neural network technique whereas they modified this tool to exploit the relations among the biomedical terms semantically.

SRL is a procedure to the idea and terms roles in a text document. The main concept is that the subject and object roles in the text document are determined based on analysis of each sentence semantically. It can be produced to the categorization of actions such as determination of “who” did “what” to “whom”, “where”, “how”, and “when”. A verb is usually used as clause predicate begins “what” took place, and other fragments of the sentence definite the other roles of the sentence (such as “when” and “who”). The main function of SRL is to define the semantic relations between a predicate and its share properties or participants, with these relationships drawn from a predefined list of potential semantic roles for the class of predicate or the predicate itself. The dataset set that mainly used in the SRL and the SENNA tools is called Proposition Bank (Propbank) corpus.

In this study, role frequency of semantic-part marking in view of the sentence-based was suggested as a new technique for idea plagiarism identification. SRL intends to identify the game plan likeness among the ideas of the reports and conceivable semantic closeness among both records. This progression in the review utilized the part marks of the ideas for the text-documents and gathered them as clusters. The cluster that was utilized as a part of this technique gave a snappy manual for capturing the associated part with the text.

B. Text Preprocessing

In this phase, the text preprocessing stage contained three sub-stages which were text chunk, and stop words withdrawal. A text chunk partitioned a text archive into sub-sentences. Several studies concentrate on text preparing strategies in various fields, incorporate intrusion detection [23]. The step of stop terms removal for erasing meaningless terms was utilized. This progression separated the critical terms from the text and disregarded the rest of the terms. This may have unfavourably influenced the comparability between texts.

A basic prepreparing includes isolating the text into important parts and is defined text chunking. Text can be separated into words, themes, or sentences. The chunking is conducted by limit recognition and isolating a text into sub-sentences. By and large, an outcry stamp (!), a question mark (?), or a period (.) is the typical signs that show a sentences limit Mikheev [24]. This study utilized the sentence based text

chunking as the initial phase in the suggested approach, where the first and suspected documents will be isolated into sentence pieces. This technique was picked on the grounds that our proposed strategy intends to contrast a speculated text and unique text in light of the sentence matching methodology.

Stop Terms are the Terms that every now and again happen in archives. They are Terms, for example, "a", "and" and "the". These terms don't provide any indication qualities or implications to the substance of the records, henceforth, they are dispensed with from the arrangement of file words [25]. The proposed strategy dispensed with all the stop terms in the documents to accelerate the system procedure. The introduced strategy utilized the list of the Buckley stop terms [26] that was utilized as a part of the SMART data recovery framework at Cornell University.

C. Term and Role Frequency

Term frequency is a one of the information retrieval process for highlighting the important terms within the text. Several page ranking techniques are working based on the term frequency to sort and rank the retrieved information based on the similarity and term frequency with the user query. Some retrieved documents can be similar and contribute more than other documents but will not be selected as relevant due to less term frequency between the query and the corpus. The idea of these documents can be identical semantically, but the similarity techniques ant not able to capture the similar pattern and meaning due to the lexical and character matching approached that was used in their techniques. This one of the main issues will be solved by the proposed method.

This research proposes a role frequency rather than term frequency to capture the idea of the documents. It was noticed that form the introduced technique the frequency of the term role can contribute more than the term frequency in some documents spatially when the people plagiarized the document idea by changing the structure of the sentences and reword the terms if text with their synonyms. The term can be either a verb or a role; either a word or phrase and the role can a labeled term.

A Verb argument structure is a useful example of semantic structure extraction and role term frequency capturing:

Martin eats the banana

eats: the verb

Martin & the banana: roles of the verb “eats”, Label: assigned to a role, Martin: subject, the banana: object.

TABLE I. ARGUMENTS SORTS AND THEIR PORTRAYALS [1]

Argument Description	Argument Type	Argument Description	Argument Type
Agent	Arg(0)	Negation Marker	NEG
Object	Arg(1)	Location	LOC
Not-fixed	Arg(2 t0 5)	Purpose	PNC
Verb	V	Modal-verb	MOD
Manner	MNR	Direction	DIR
Time	TMP	Exit	EXIT
General purpose	ADV	Discourse connectives	DIS

The meaning of the arguments types illustrates in Table I.

Table I demonstrates the sorts of arguments that were utilized as a part of the analyses and their depiction or significance.

Analysis of the text document based on the term role of each sentence in the document is one of the main steps of the introduced method. Initially, the terms roles are extracting using the semantic role labeling NLP technique. The function of the SRL is to exploit roles and roles for all terms inside the sentence. The SRL used the analysis of the paragraph and sentence based on the sentence predicate (Subject, Verb, and Object) to define each the role of each term in the sentence and paragraph semantically using the PropBank dataset [17]. The verb argument plays an important role of the proposed method by extracting the semantic structure and the term frequency. The role term frequency (rtf) defined as the occurrence times of role term r in verb role structures of paragraphs. The role term (r) can be a word or phrase. The rtf of term (r) in document (d) can have various rtf parameters in a different paragraph in document (d), these parameters formulated as:

$$rtf = \frac{\sum_{n=1}^{sn} rtf_n}{sn} \quad (1)$$

Where sn is the number of paragraphs or sentences that hold role term r in text document (d).

IV. IDEA PLAGIARISM DETECTION

Theft and copy of ideas are considered one of the intelligent types of text plagiarism, especially in the scientific text and articles. The essential idea from the originals text is manipulated, exploited and represented in the suspicious text as a novel or new. Within a text, the idea can be presented in sentences, paragraphs, or phrases. The method developed extracts semantic ideas within a text based on term role frequency. The plagiarism detection is conducted based on two levels; Sentence level detection, and paragraph level detection. In the sentence detection level, the SRL explores sentences in the document based on the role of each term inside the sentences. Then, the frequency of extracted roles calculated and considered as important semantically. These roles frequencies are then utilized in capturing the similarity in both sentence level and paragraph level. On the other hand, the paragraph level detection is conducted based on the semantic meaning of the group of the roles terms frequencies that were extracted from the sentences level detection. Fig. 2 illustrates the idea extraction from the document.

The idea of the sentences extracts by focusing on the terms that have contributed more than other terms in the each paragraph. The term that has one frequency will be ignored and the terms that have a more frequency will be considered as significant terms. Finally, the ideas will be categorized into roles and arguments. The main idea of the paragraph is equal the group of sentences ideas in that paragraph. Typically, the main idea of the original and suspected documents is equal the group of paragraph ideas in that document and the title idea accordingly. An example of how the idea extracted based on term role frequency demonstrates in following original and suspected sentences:

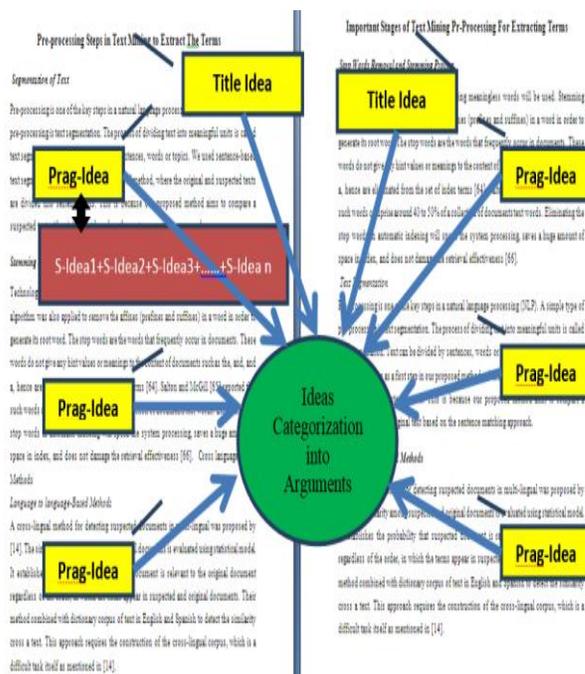


Fig. 2. Idea Extraction from the Document.

Example 1 supported by Shehata and et al. [27]:

Original sentence:

Texas and Australia researchers have created industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles.

Suspected sentence:

The industry-ready sheets that were created by Malaysia and China researchers from resources made from nanotubes that could lead to the development of artificial muscles.

We noted from the example some important pints that should be the focus:

- Each Term has one frequency,
- What are the Important Terms?
- What is the term that has contributed more than the other terms?

The SRL analysis of calculating rtf of example 1 demonstrates as:

Original sentence:

First, the SRL will be employed to extract the verbs the sentence. In this example, three target words (verbs) is extracted as a verb argument structures:

- created
- made
- Lead

Fig. 3, 4, 5 and 6 illustrates the SRL analysis of the original sentence verbs [18] [33].

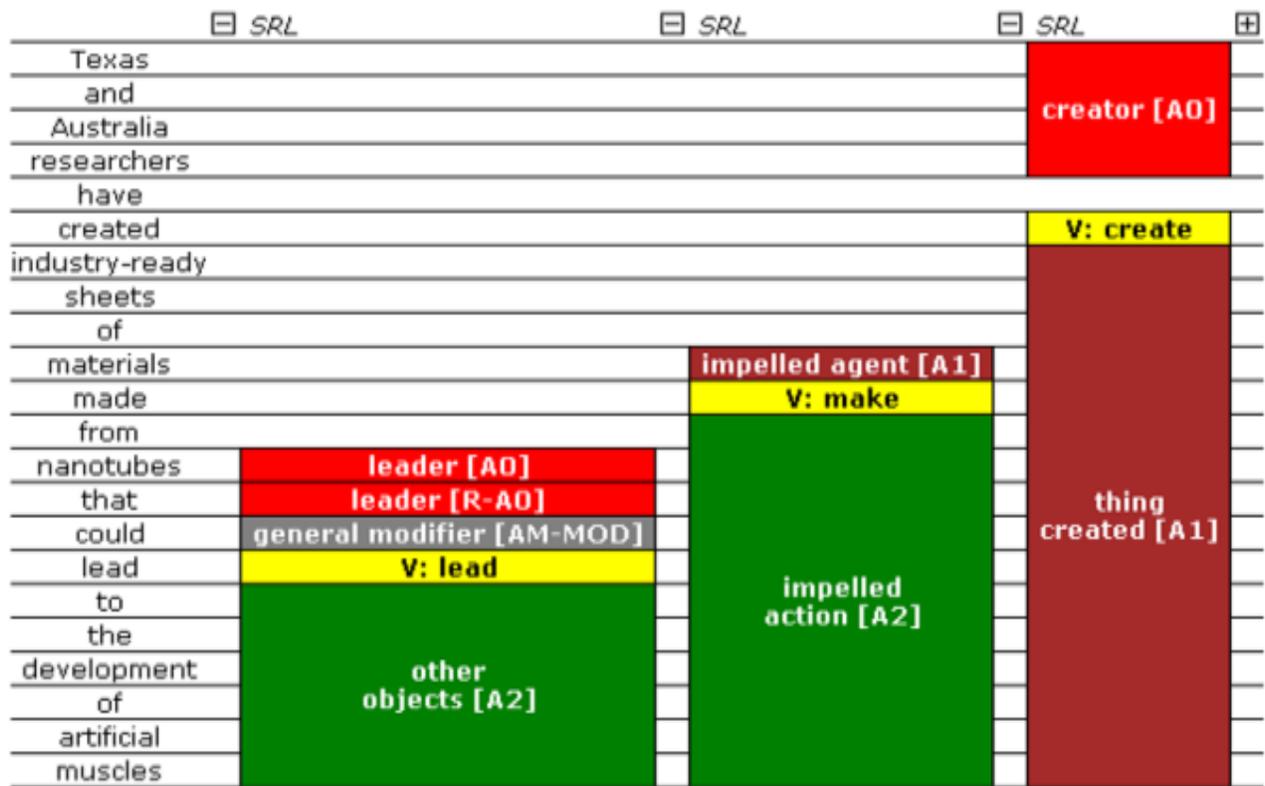


Fig. 3. Analysis the Original Sentence using SRL.

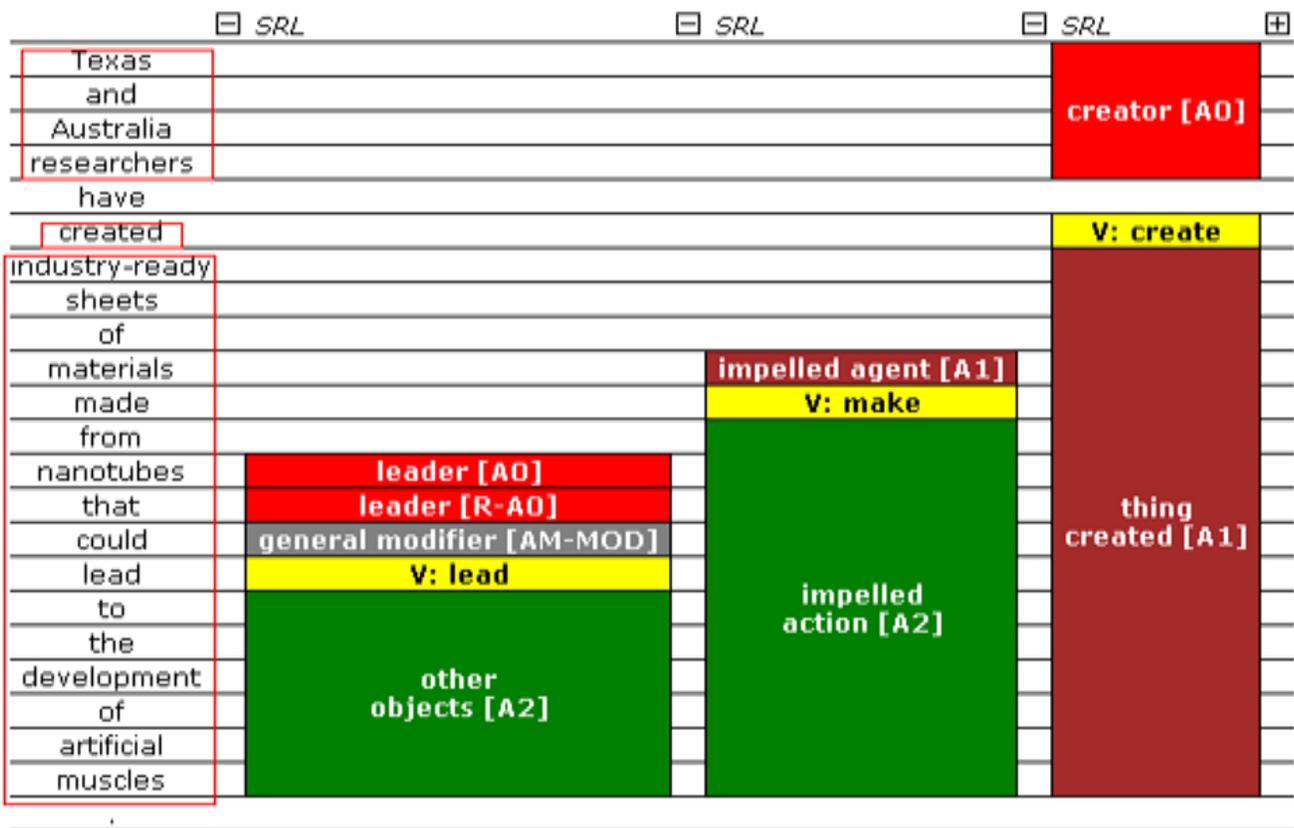


Fig. 4. Calculating rtf of the Created Verb Argument Structure.

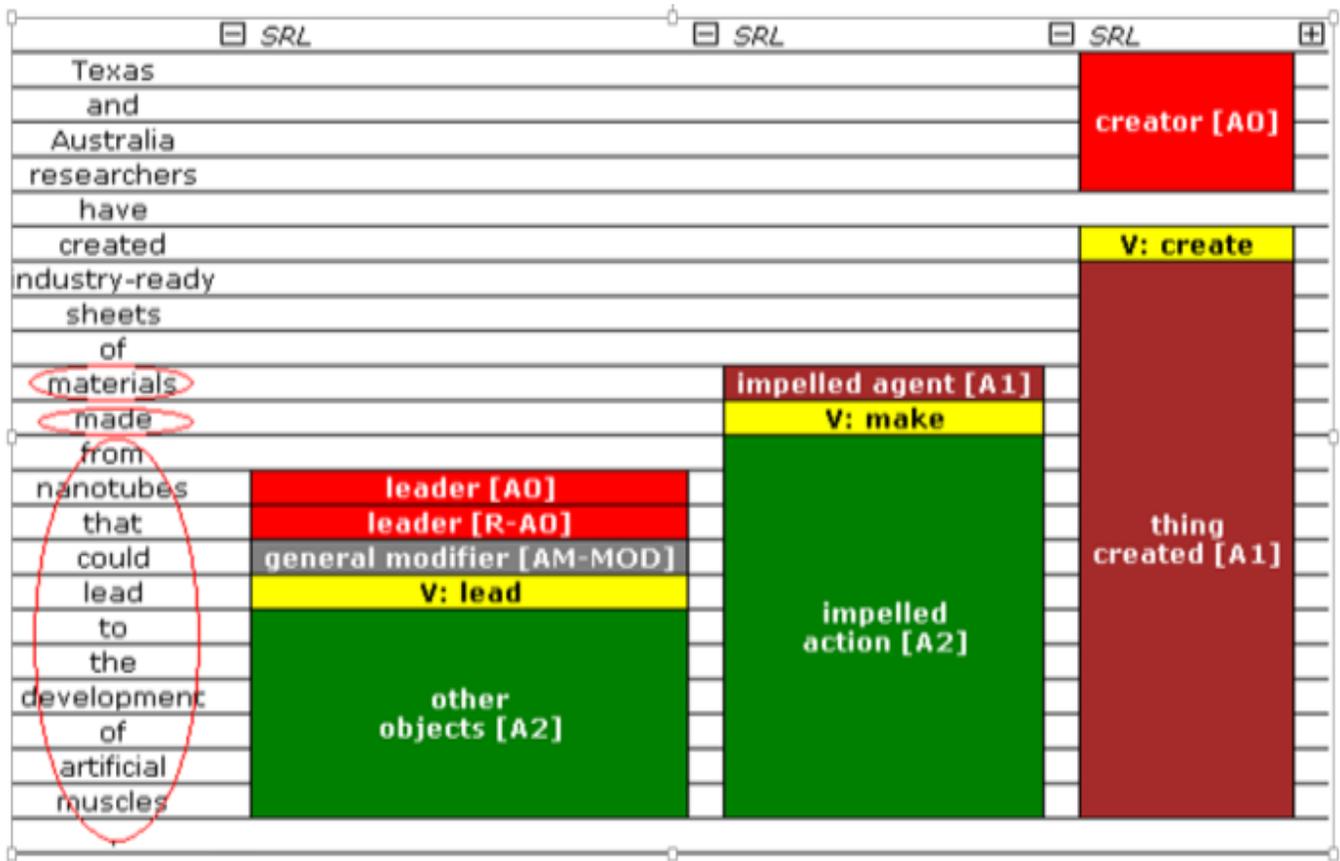


Fig. 5. Calculating rtf of the *Made* Verb Argument Structure.

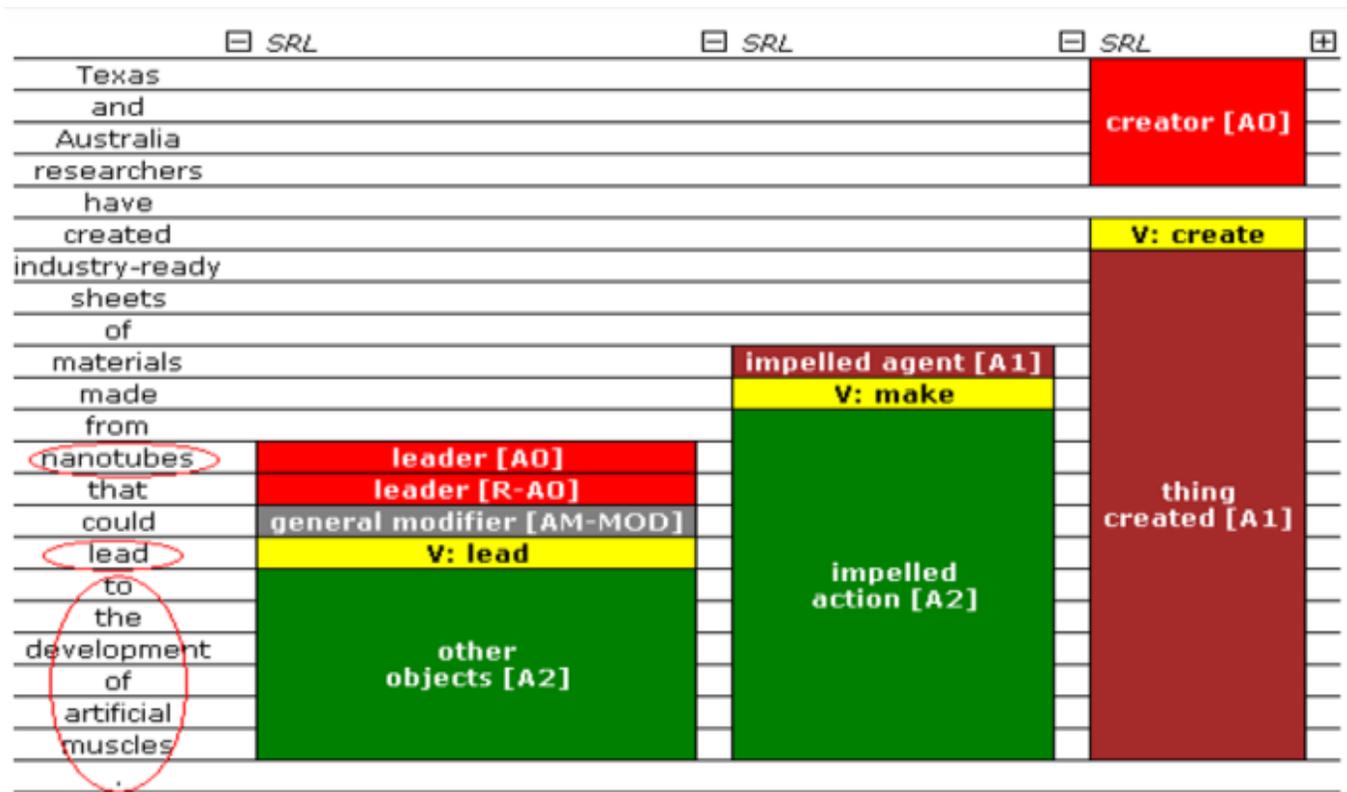


Fig. 6. Calculating rtf of the *Lead* Verb Argument Structure.

TABLE II. TERM ROLE FREQUENCY OF THE ORIGINAL SENTENCE

Role NO	Role-Term	Role Frequency	Role NO	Individual-Role-Term	Role term Frequency
1	-Texas-and-Australia-Researchers	1	1	- Texas	1
				-Australia	1
				Researchers-	1
2	Created-	1	2		1
3	-Industry-rely-sheets materials-made-nanotubes-lead development artificial-muscles	1	3	-Industry	1
				-rely	1
				-sheets	1
4	Materials-	1	4		1
5	nanotubes-lead-development artificial-muscles	1	5	Development-	3
				Artificial-	3
6	Nanotubes-	3	6	-	
7	-lead	3	7	-	
8	- development artificial-muscles	3	8	-muscles	3

Argument (1) is (Texas and Australia researchers), **the TARGET is made, and** Argument (2) *is* (industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles).

Argument (1) is (nanotubes), the TARGET is **lead**, and Argument (2) *is* (to the development of artificial muscles). The terms roles frequency at the first chunk of the original sentence in this example is calculating in both, phrase term and individual role term. Table II illustrates the term role frequency of the original sentence.

The Idea of the original sentence is “development artificial muscles” because the terms (development, artificial, and muscles) has a high role frequency and contribute more than other terms.

Suspected sentence:

The industry-ready sheets that were created by Malaysia and China researchers from resources made from nanotubes that could lead to the development of artificial muscles.

Three target words (verbs) for is also extracted from the suspected sentence. The extracted verb argument structures of this example are:

- created
- made
- Lead

We employed the same process that was used in the original sentence by extracting the argument structure of the verb **mad**, and **lead** is:

The verb (Made):

- Argument (1) is (materials), **the TARGET is made, and** Argument (2) *is* (nanotubes that could lead to the development of artificial muscles).

The verb (Lead):

Argument (1) is (nanotubes), **the TARGET is lead, and** Argument (2) *is* (to the development of artificial muscles). Table III illustrates the term role frequency of the suspected sentence

The Idea of the suspected sentence is “development artificial muscles” because the terms (development, artificial, and muscles) has a high role frequency and contribute more than other terms. We noted from calculating the rtf that it is possible two frequent terms have the same occurrence in their document, but one pays more to the meaning of its sentence than the counter one. This can occur when the people plagiarize the main idea of the documents with changing the structure of the documents or performing semantic plagiarism. This type called idea plagiarism. Additionally, we noted that the rtf could assist for capturing the similarity between the documents by extracting the main document idea.

TABLE III. TERM ROLE FREQUENCY OF THE SUSPECTED SENTENCE

Role -No	Role-Term	Role-Freque ncy	Role -No	Individual-Role-Term	Role-term-Freque ncy
1	Industry-rely-sheets	1	1	-Industry	1
				-rely	1
				-sheets	1
2	Created-	1	2		-
3	-Malaysia-China-researchers-resources-made-from-nanotubes-lead-development artificial-muscles.-	1	3	-Malaysia	1
				-China	1
				-resources	1
4	-resources	2	4	-	-
5	nanotubes-lead-development artificial-muscles	2	5	Development	3
				Artificial-	3
6	Nanotubes-	3	-6	-	-
7	-lead	3	7	-	-
8	- development artificial-muscles	3	8	-muscles	3

V. EXPERIMENTAL DESIGN

A. Corpus and Dataset

The CS11 dataset comprises 100 human short answer questions samples of plagiarized text collected by Clough and Stevenson [2]. It offers cases of plagiarized short texts made in various plagiarism levels. The benefit of the CS11 dataset is that it is simulated and developed by a human; wherein the behaviour situation of plagiarized peoples is natural not artificial. The dataset involves of a 100 text 95 suspected short texts and 5 articles collected from Wikipedia website as the original documents. Non-native and native scholars to response five questions interrelated to the original articles wrote the suspicious texts. The responses were excepting for non-plagiarized samples based on the original documents with varied similarity, as well as the instructions are specified by the dataset inventers. The average terms in the short texts were among (200 - 300). Around 57 cases were noticeable: out of which 19 is heavy revision, 19 were nearby copy, and additional 19 marked as light revision samples; while the resting 38 cases were free plagiarized texts. The various kinds of suspected texts are defined as:

- 1) *Near copy*: this type focuses on a copy and paste from the original text;
- 2) *Light revision*: minor alteration of the original documents by substituting terms with their synonyms and giving a little linguistic modifications;
- 3) *Heavy revision*: rewriting and major alteration with rephrasing and restructuring in original documents;
- 4) *Non-plagiarism*: revised texts without any alteration in the original documents based on contributors' own terms.

B. Experimental Design

The experiments of the proposed method used the Clough09 Corpus for detecting the plagiarized idea. This is because of semantic characteristics imitation for plagiarism samples such as text rephrasing. To examine the proposed methods in, we utilized the CS11 dataset that was designed because of the PAN-PC dataset limitations. The limitations are that the common of the plagiarized samples were produced artificially. The experiments examined the amount of detecting plagiarized sentences from the original documents based on the Clough09 plagiarism Corpus. The suggested technique analysis the source and suspected documents based on the SRL. The analyzed sentences are then used to calculate the rtf from each sentence in the corpus. The idea extracted based on the rtf from the sources and suspected documents. The similarity between the ideas in the both documents is calculates based on the proposed role-based similarity metrics. The rft-based similarity metrics is adopted to detect the matching between the texts based on sentences and documents levels. To adopt the rtf-based similarity a role term-based analysis will be structured and formulated. This analysis considers two main issues; Similarity measure based on rtf, tf and df of matched role terms, and matched role terms (role terms that exist in two or more documents).

The Role Term-Based Similarity measure between two-text documents d_1, d_2 , used factors:

- m : number of matching role terms between d_1 and d_2
- sn : the gross number of sentences hold similar role term r_i in every text document d
- li : length of each role term in the verb role construction in very text document d
- Lvi : length of very verb role construction which holds the similar role term r_i .
- N : gross number of text documents in the dataset

The similarity measure between two document d_1, d_2 is calculated as:

$$sim_c(d_1, d_2) = \sum_{i=1}^n \max\left(\frac{li_1}{li_{u12}}, \frac{li_2}{li_{u12}}\right) * weight_{t_{i1}} * weight_{t_{i2}} \quad (2)$$

$$weight_i = (tfweight_i + cfweight_i) * \log\left(\frac{N}{df_i}\right) \quad (3)$$

$tfweight_i$ =document level; $tfweight_i$ = Sentence level

Where $weight_i$ denotes the weight of the role term i in text document d .

The $tfweight$ and $rtfweight$ are normalized by length of document vector

$$tfweight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn} (tf_{ij})^2}} \quad (4)$$

$$ctfweight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn} (ctf_{ij})^2}} \quad (5)$$

Where cn : the gross number of role terms which has a terms frequency values in text document d .

C. Performance Measures

This unit deliberates evaluation measures of plagiarism detection methods. The Precision and Recall factor is a common evaluation measure that normally used to assist the plagiarism detection. Potthast et al., [28-30] suggested a macro-averaged and a micro-averaged variant. The granularity or F-measure factor is an additional significant measure that was utilized in plagiarism detection performance [31][30][31][31][31](Potthast et al., 2010b), (Potthast et al., 2010b), (Potthast et al., 2010b), and (Potthast et al., 2010b). We use the micro-averaged Recall and Precision for evaluating our proposed method. The Precision and Recall of R under S are identified as follows:

$$Precision_{micro}(S, R) = \frac{|U_{(s,r) \in (S \times R)}(S \cap R)|}{|U_{r \in R} r|} \quad (6)$$

Where, S and R present sets of plagiarized samples and detections, s denote plagiarized passage in a plagiarized documents, r denote associates a supposedly plagiarized passage in documents.

$$Recall_{micro}(S, R) = \frac{|U_{(s,r) \in (S \times R)}(S \cap R)|}{|U_{s \in S} s|} \quad (7)$$

Where

$$S \cap R = \begin{cases} s \cap r & \text{if } r \text{ detect } s \\ \emptyset & \text{Otherwise} \end{cases} \quad (8)$$

The granularity is the harmonic mean of recall and precision and computed based on the following formula:

$$\text{Granularity} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (9)$$

VI. RESULTS AND DISCUSSION

The results of the matching computation in the precision, recall, and granularity are specified in Fig. 7, 8, and 9 for different plagiarism classes Heavy, Light, and, Cut-and-paste respectively. The results of our proposed method are compared with other method reported by Chong [15]. It utilizing Naïve Bayes classifier with an arrangement of all elements, best components, and Ferret Baseline method [32]. These techniques were talked about before in Section 2. We select these techniques for correlation since it utilizes the CS11 human short answers question corpus. The aftereffects of the correlations show additionally in Fig. 7, 8, and 9.

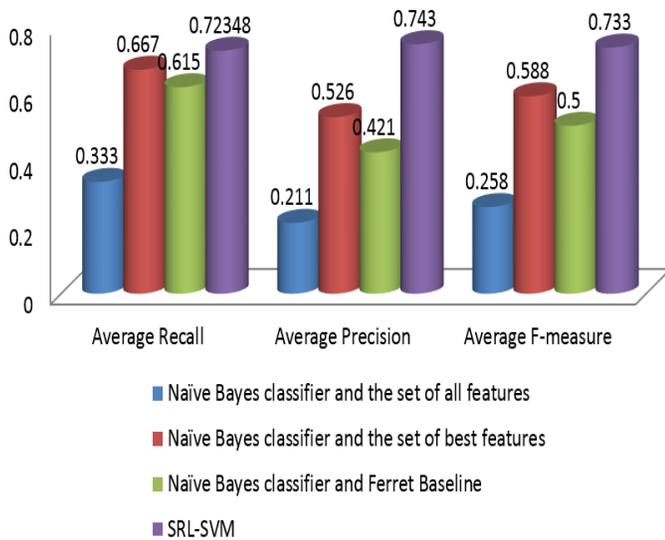


Fig. 7. CS11 Heavy Plagiarized Samples.

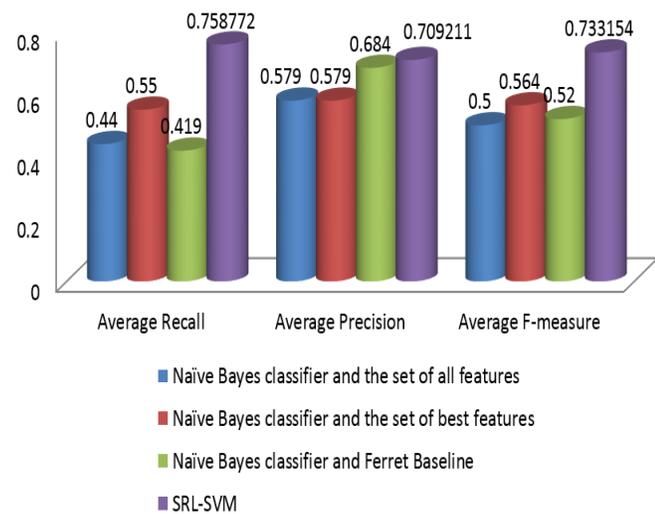


Fig. 8. CS11 Light Plagiarized Samples.

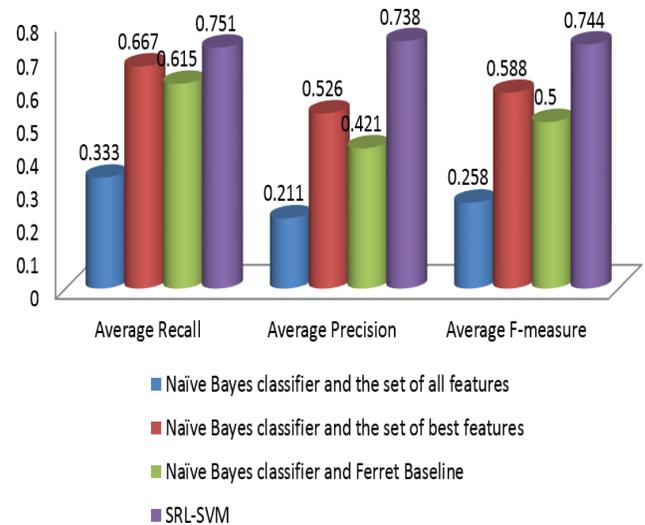


Fig. 9. CS11 Cut-and-Paste Plagiarized Samples.

Fig. 7 showed the correlation comes about between the suggested technique and different strategies based on Heavy copyright infringement class. We noticed that the suggested strategy accomplished high scores in term of Recall (0.723), Precision (0.743), and Granularity (0.733).

Fig. 8 showed the examination comes about between the suggested technique and different strategies based on light unoriginality class. We noticed that the suggested technique accomplished best scores in term of Recall (0.758), Precision (0.709), and Granularity (0.733).

Fig. 9 shown the cut-and-paste results in term of Recall, Precision, and Granularity with (0.751), (0.738), and (0.744) respectively.

The experimental output showed that the idea similarity crosses the CS11 corpus proved that the general performance in the precision, recall, and Granularity are achieved better results for capturing the main idea between the source texts and suspicion texts in the dataset. The proposed method tested with the different types of plagiarism in the CS11 corpus such as heavy, light, and copy-and-paste. Through the results, we observed that the suggested technique obtained good results compared with Naïve Bayes method with an arrangement of all elements, best components, and Ferret Baseline technique [32].

VII. CONCLUSION AND FUTURE WORK

In this research, an idea plagiarism detection system using term role frequency is suggested and explained. The suggested method analyzed and compared the idea of the text based on role based-similarity and the frequency for each term in a text. It is possible that two frequent words have the identical occurrence in their document, but one pays more to the semantic of its sentence than the counter one. These documents can also have the same idea, but the main structure of the idea presentation totally differs spatially in the structure and semantic meaning. Semantic Role Labeling obtained significant benefits when it originated to produce meaning roles for every sentence individually. The utilization is to

detect the semantic matching among the passages. The main contributions and idea of the documents can be extracted by calculating the important roles using role term frequency. The results of the similarity performed and calculated across the CS11 corpus and proved that the general performance of the suggested method is succeeded to capture the main idea between the source documents and suspicion documents in the dataset. The proposed method examined with samples of plagiarized text in diverse levels of plagiarism such as cut and paste, minor modification (light), and major rephrasing in source texts (Heavy). The results of the term role frequency discovered the benefits of the role-based term similarity for detecting the plagiarized idea between the source and suspected documents. In the future, the suggested method will be combining with optimization technique to improve the performance results. Another dataset using PAN-10 to PAN-12 will be tested and examined.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (G: 466-830-1439). The author, therefore, gratefully acknowledges the technical and financial support from the DSR.

REFERENCES

- [1] A. H. Osman, N. Salim, M. S. Binwahlan, R. Alteeb, and A. Abuobieda, "An improved plagiarism detection scheme based on semantic role labeling," *Applied Soft Computing*, vol. 12, pp. 1493-1502, 2012.
- [2] P. Clough and M. Stevenson, "Developing a corpus of plagiarised short answers," *Lang. Resour. Eval.*, vol. 45, pp. 5-24, 2011.
- [3] S. Alzahrani, V. Palade, N. Salim, and A. Abraham, "Using structural information and citation evidence to detect significant plagiarism cases in scientific publications," *Journal of the American Society for Information Science and Technology*, pp. n/a-n/a, 2011.
- [4] M. Hermann, K. Frank, and Z. Bilal, "Plagiarism - A Survey," *Journal of Universal Computer Science*, vol. 12, pp. 1050-1084., 2006.
- [5] K. Vani and D. Gupta, "Detection of idea plagiarism using syntax-Semantic concept extractions with genetic algorithm," *Expert systems with applications*, vol. 73, pp. 11-26, 2017.
- [6] D. Gupta, "Study on Extrinsic Text Plagiarism Detection Techniques and Tools," *Journal of Engineering Science & Technology Review*, vol. 9, 2016.
- [7] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, IEEE Transactions on, vol. PP, pp. 1-1, 2011.
- [8] J. D. Velásquez, Y. Covacevich, F. Molina, E. Marrese-Taylor, C. Rodríguez, and F. Bravo-Marquez, "DOCODE 3.0 (DOCUMENT COPY DETECTOR): A system for plagiarism detection by applying an information fusion process from multiple documental data sources," *Information Fusion*, vol. 27, pp. 64-75, 2016.
- [9] D. Weber-Wulff, *False feathers: A perspective on academic plagiarism*: Springer Science & Business, 2014.
- [10] M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso, "Overview of the 1st International Competition on Plagiarism Detection," in *PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection*, 2009, pp. 1-9.
- [11] A. H. Osman, N. Salim, and A. Abuobieda, "Survey of text plagiarism detection," *Computer Engineering and Applications Journal (ComEngApp)*, vol. 1, pp. 37-45, 2012.
- [12] Y. Palkovskii, A. Belov, and I. Muzika, "Exploring Fingerprinting as External Plagiarism Detection Method," 2010.
- [13] Y. Palkovskii, A. Belov, and I. Muzyka, "Using WordNet-based semantic similarity measurement in External Plagiarism Detection," in *CLEF (Notebook Papers/LABs/Workshops)*, Amsterdam, The Netherlands, 2011.
- [14] G. Parth, R. Sameer, and P. Majumdar, "External Plagiarism Detection: N-Gram Approach using Named Entity Recognizer," presented at the *CLEF (Notebook Papers/LABs/Workshops) 2010*.
- [15] M. Chong, L. Specia, and R. Mitkov, "Using Natural Language Processing for Automatic Detection of Plagiarism," *Proceedings of the 4th International Plagiarism*, 2010.
- [16] C. J. Fillmore, "The case for case. In Emmon Bach and Robert T," *Universals in Linguistic Theory*. Holt, Rinehart, and Winston, New York, pp. 1-210, 1968.
- [17] Martha Palmer, Daniel Gildea, and Paul Kingsbury, "The Proposition Bank: An Annotated Corpus of Semantic Roles," *Comput. Linguist.*, vol. 31, pp. 71-106, 2005.
- [18] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet Project," presented at the *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, Montreal, Quebec, Canada, 1998.
- [19] Pradhan, S. Sameer, H. Wayne, K. H. Ward, H. M. James, and Dan Jurafsky, "Shallow semantic parsing using support vector machines," in *Proceedings of NAACL-HLT 2004*, pp. 233-240, 2004.
- [20] M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth, "Using predicate-argument structures for information extraction," presented at the *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, Sapporo, Japan, 2003.
- [21] Xue, Nianwen, and Martha Palmer, "Calibrating features for semantic role labeling," in *Proceedings of EMNLP 2004*, pp. 88-94, 2004.
- [22] T. Barnickel, J. Weston, R. Collobert, H. Mewes, and V. Stumpflen, "Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts," *International Journal of Information Technology & Decision Making*, vol. 4, p. e6393, 2009.
- [23] A. Sharma, A. K. Pujari, and K. K. Paliwal, "Intrusion detection using text processing techniques with a kernel based similarity measure," *Computers and Security*, vol. 26, pp. 488-495, 2007.
- [24] A. Mikheev, "Document centered approach to text normalization," in *Proceedings of SIGIR*, pp. 136-143 2000.
- [25] C. Rijsbergen and J. Van, "A New Theoretical Framework for Information Retrieval.," 1979.
- [26] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic query expansion using SMART: TREC 3," *NIST SPECIAL PUBLICATION SP*, pp. 69-69, 1995.
- [27] S. Shehata, F. Karray, and M. Kamel, "An efficient concept-based mining model for enhancing text clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1360-1371, 2010.
- [28] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, and P. Rosso, "Overview of the 2nd international competition on plagiarism detection," *Notebook Papers of CLEF*, vol. 10, 2010.
- [29] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, and P. Rosso, "Overview of the 3rd international competition on plagiarism detection," *Notebook Papers of CLEF 11*, vol. 10, 2011.
- [30] B. Stein, M. Potthast, P. Rosso, A. Barrón-Cedeno, E. Stamatatos, and M. Koppel, "Fourth international workshop on uncovering plagiarism, authorship, and social software misuse," in *ACM SIGIR Forum*, 2011, pp. 45-48.
- [31] M. Potthast, B. Stein, A. Barr, #243, n-Cede, #241, et al., "An evaluation framework for plagiarism detection," presented at the *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, China, 2010.
- [32] C. Lyon, R. Barrett, and J. Malcolm, "A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector," *Plagiarism: Prevention, Practice and Policies Conference Newcastle, UK*. 2004, 2004.

Impact of Security in QoS Signaling in NGN: Registration Study

RAOUYANE Brahim¹

N&DP Team, IT&NT Laboratory,
Faculty of Sciences, Ain Chock, Casablanca, Morocco

BELMEKKI Elmostafa², KHAIRI sara³,

BELLAFKIH mostafa⁴

RAISS Team, INPT
Rabat, Morocco

Abstract—New generation networks (NGN) use an IP base to transmit their services as well as voice, video and other services. The IP Multimedia Subsystem (IMS) which represents the network core, allowed controls and accesses into various services through a set of signalling protocols, the most common of which is Session Initiation Protocol (SIP). After securing the most vulnerable interfaces in the core of the NGN: IMS architecture. The idea is to improve QoS in SIP signalling, especially in authentication and registration that represent the first step to access. The proposed approach is used as encryption asymmetry in the SIP registration process and study the performance of the system in terms of QoS parameters.

Keywords—Quality of Service (QoS); Security; New Generation Network (NGN); IP Multimedia Subsystem (IMS); Session Initiation Protocol (SIP)

I. INTRODUCTION

The Next Generation Network (NGN) enables [1] the deployment of independent access services over fixed and mobile networks with agnostic convergence. NGN is based on packet switching and uses IP to transport different types of traffic (voice, video, data and signalling). The specifications were agreed that the EPC (Evolved Packet Core) would no longer have a circuit-switched domain and that the EPC should be an evolution of the packet switching architecture used in GPRS / UMTS. Indeed, this decision had consequences on the architecture itself but also on the way services were provided. Security in NGN is important, the main goal is to choose a sensitive scenario to deal with. According to security analysis and modelling, NGN can be divided into 3 boxes or boxes with their communication interfaces. The risk analysis with EBIOS (Expression of Needs and Identification of Security Objectives) designates three boxes: Client, IMS, and Server, the communication between these boxes is made by standard protocols signalling or service according to the customer need. The most commonly used services with the IMS is recording, calls, videos, messages. etc. The primary service performed by each connection is the registration, this operation is important also sensitive to faults that are related to the use or inherited from the packet switching network and others attached to the SIP signalling protocol. Our approach to remedy this registration problem is reinforced security in SIP before using the regular methods of IP (SSH, TLS). The approach is based on the Register scenario study as well as SDL (Specification and Description Language) modelling and finally demonstrates its theoretical and practical reliability in a test network.

Security issues in the IMS network is an important challenge as it includes a wide variety of services, protocols and components. This complexity enhances the number of vulnerabilities and risk for the IMS users and the ISP (Internet Service Provider). Some of these vulnerabilities are inherent on one hand to protocols and services used and others are induced by the context of the IMS like users mobility. On the other hand, QoS is also big challenge in any IMS network as this network is designed to offer time sensitive application like video, videoconferencing and so on. The main idea in this paper is to secure IMS services and evaluate the impact on QoS as well as [2][3].

In this work we will first present the IMS network architecture and we propose a state of the art of the IMS network. Second, we present our approach to secure the SIP registration after having identifying interfaces and sensitive entities in the architecture. Finally, we will analyse experimentally the operational of primordial protocols as SIP proposed compared to security standards to highlight all associated loopholes.

II. PRESENTATION OF THE NGN ARCHITECTURE

As Fig. 1, the 4G / LTE (Long Term Evolution) [4] network benefits from a large flow evolution and thus services that have a direct impact on topology and structure. A user (UE) connects via eNodeB, EPC and the IP Multimedia Subsystem (IMS). The EPC is combined with E-UTRAN (Evolved Terrestrial Radio Access Network) it is the communication part of a mobile network, these composite entities to form Evolved Packet System (EPS). The EPC contains the following components: Serving Gateway (S-GW), Mobile Management Entity (MME), Policy Control and RulesFunction (PCRF), and PDN Gateway (P-GW).

The integration of new features such as SDN and virtualization into the current EPC is a complex task that involves carefully evaluating 3rd Generation Partnership Project (3GPP) standardizations. The most important challenge is to preserve LTE (Long-Term Evolution) functionality in a new, flexible and centralized EPC architecture based on new features. The proposed architecture is to redefine the main procedures for control and data plans by relying on new techniques such as SDN and virtualization functions. First, firstly, the challenges of this new architecture are discussed and the proposed solutions are presented. Secondly, the possible improvements are studied in terms of flexibility, complexity

and technology-based performance that could possibly optimize the design of the proposed system.

The proposed architecture aims at slightly modifying the existing 3GPP architecture in order to integrate existing core components, especially in the EPC and in the transport layer in order to integrate SDN [5] and OpenFlow [6] controller, also in the Service layer in order to perform changes to improve the performance of certain services with the principle of virtualization by complying with SLA (Service Level Agreement) constraints. The main basic interfaces: the S1-MME, S1-U, S6a and Gx functionalities are maintained, as well as 3GPP intra-3GPP authentication, authorization and mobility management by Mobility Management Entity (MME). The current Service Gateway (SGW) / Packet Data Network Gateway (PGW) selection mechanism based on the Domain Name System (DNS) has been changed. The MME queries the OpenFlow controller through the NorthBound interface (representation state transfer API) that can install transfer rules in OpenFlow switches.

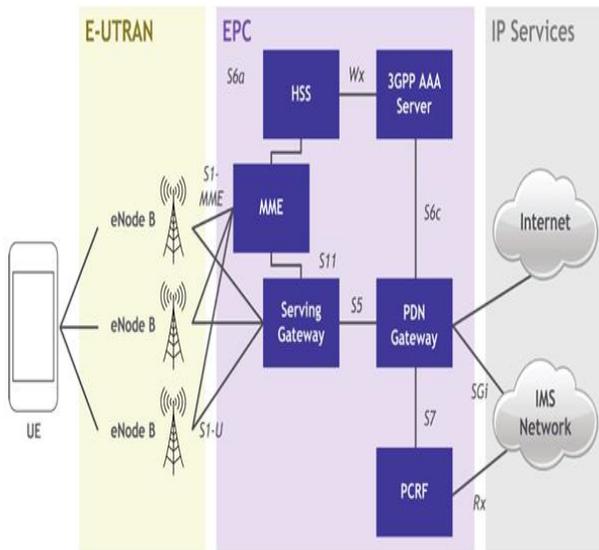


Fig. 1. IMS avec LTE-Evolved Packet Core [5].

A. SIP Security in NGN

The proposed new approach begins with an analysis of the IMS architecture to secure the communication interfaces. To this end, a study is made with the useful EBIOS to explore the main network entities as well as their communications interfaces (Fig. 2) [7].

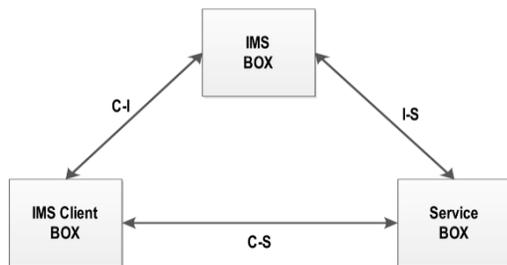


Fig. 2. Modelling of NGN / IMS Entities and their Communication Interfaces.

The resulting model consists of three entities:

- **IMS-Client BOX:** Represents the end user connected to the IMS to access network services. The IMS-Client can access it either through a packet or circuit network.
- **IMS-BOX:** represents the core of the IMS network with its different internal components with a direct and secure connection. The box represents an abstraction of the details of the communications and the different operations that take place at the heart.
- **Service-BOX:** Represents the service platforms provided by IMS for clients.
- The components communicate via 3 interfaces:
- **C-I interface:** Between the client and the IMS-BOX. The interface transports all signalling and access control traffic between the client and the IMS network. the most used protocol is SIP.
- **I-S interface:** An interface between the Service-BOX service platforms and the IMS-BOX IMS. It comprises, on the one hand, the signals that make it possible to verify that the service platform is authenticated and authorized to provide the service via the IMS network and, on the other hand, that the communication between a server and a client by the request for service access to a service. The I-S interface is based on the SIP protocol.
- **C-S interface:** Between the client and the service platforms, the interface contains the traffic of the service requested by the client. Multi-form traffic is VoIP communication, video conferencing, etc. The protocols used on this interface are related to the types of services.

The study proved that the communication between IMS core entities such as CSCF and HSS is very secure, since the communication link is direct as well as the security mechanism uses certificates with Diameter protocol, these assets ensure confidentiality, integrity and authentication. Nevertheless, the traffic or the flow of information, which circulates between the client and the IMS core or between the application server and the IMS core, remain vulnerable since they cross different others network entities. These communication interfaces have a critical degree of severity that classifies the risks on these interfaces as an operator security violation that must be handled first, in relation to the services that it may be SIP, HTTP, RTP, FTP traffic, or others [7].

TABLE I. PROTOCOLS AND COMMUNICATIONS INTERFACES IN NGN FOR VOLTE

Interfaces	Protocols	Security Mechanism
C-I	SIP DHCP DNS	IPsec TLS
C-S	RTP SIP	sRTP IPsec
I-S	SIP	IPsec TLS

3GPP's proposals for security in NGN / IMS architecture have difficulties and limitations, hence the need for convergence to other solutions to improve and enhance security in the three communication interfaces. The first step is to identify and analyse the traffic passing between the three components. Table I provides an overview of the set of signalling protocols and services in NGN as well as the security mechanisms recommended by the 3GPP and RFC standards.

The signalling protocols always carry out the services, for this it is necessary to secure these protocols upstream and downstream. SIP is the primary signalling protocol that will be our goal in the IMS context. Indeed, SIP is a signalling protocol that specifies the exchange of information to manage multimedia sessions in the IMS. The protocol describes the power to establish, modify and terminate a multimedia session [8].

The IMS benefits from RFC specifications and uses mechanisms to provide communication between these Client/Server/Proxy entities. The text-based SIP and uses HTTPDigest for authentication and user registration, also to secure TLS, S/MIME and IPsec signalling. On the other hand, these solutions guarantee a security on a domain or on a link but not on the whole of an end-to-end communication [8].

Before accessing the various services of the IMS, it is necessary to go through the first step which is essential authentication. The operation of authenticating a user is based on a simple challenge / response that contain several risks [9].

The TLS or S / MIME solutions guarantee security but require an intervention with certificates, which is not possible in the IMS infrastructures. As well as HTTPDigest remains simple, to the opposition of specifications of 3G with AKA [10] which impose that an authentication must be mutual systematically. Our contribution is to strengthen authentication at the signalling level either by decreasing the vulnerabilities of simple authentication, or by offering another form of mutual authentication with HTTP Digest [11].

B. Approach to securing SIP Authentication

The problem exists in IMS resides in the REGISTER method of the SIP protocol especially in the sensitive parts in the authentication messages, so our objective is to reinforce the mutual security between the two communicating parties without modifying the communication via the SIP protocol as well as infrastructural interoperability, while minimizing the impact on QoS. The proposed solution makes it possible to hide the sensitive fields in the SIP messages, the asymmetric encryption guarantees mutual reliable communication between the two parties (Client / Server). A partial or radical change in an existing protocol requires behavioral and static modelling to keep the properties of it. The modeling is followed by a verification test to ensure no changes in the content and SIP message only the content.

Firstly, our objective is to propose a solution to reinforce the security with a mutual authentication, what follows is to integrate the solution in the SIP protocol, secondly it is the integration in the NGN network and finally to measure the

impact on the QoS. Integration is a difficult operation considering its composition of a set of processes resulting from modeling followed by a formal validation with SDL (Specification Description Language) [12], and also by a behavioral validation with the use of the MSC (Message Sequence Chart) to validate the interactions of approach [13], while respecting the IMS network components.

The user must be subscribed to an IMS network before starting any service offered by the network providers. As much as a control subsystem, IMS following an internal communication registration procedure between CSCF and HSS, and with the EU external user.

By focusing only on the signalling exchanged with UE, since the internal communications are direct and secure by a kernel. Communication between the EU and the IMS takes into account 4 messages (Fig. 3):

- 1) A Registration Request with Register Is Sent Contains EU Identity.
- 2) The IMS responds with a 401Unauthorized message with a random nonce value.
- 3) The client sends a response that contains Response, after a key calculation.
- 4) The IMS server responds with OK 200 if the answer is correct.

The method used in the registration is challenge / response authentication with HTTPDigest. In the challenge phase, a sensitive field "WWW-Authenticate is clear in the form of a "NONCE" [13]. Then, the UE generates a response based on the previous information in two "Authorization" fields with a response value in the SIP message in the form:

$$response = H (username || realm || password) || ness || H (METHOD || Request-URI) \quad (1)$$

A simple catch can expose SIP messages that are text-based and clearly accessible, sensitive information such as nonce values, and the response generated by the client during recording appear clearly during the communication. Indeed, the knowledge of these values can generate dictionary type attacks to easily calculate the shared secret value between UE and IMS. Therefore, it is necessary to secure the communication between EU and IMS on the one hand and on the other hand to strengthen the registration between the two parties. Our idea is to generate a significant value of nonce instead of a random value. The value of nonce generated depends on the value CallID, realm, URI, secret key and Timing [14].

The approach is without change in the procedure, it adds additional functions next to Server and SIP Client (encryption / decryption).

The reinforcement scenario is illustrated in Fig. 4. This scenario, which consists of different phases, is the same as the old one.

The main idea is the generation of the new value of "nonce" with a significant value. The generated nonce is random and invisible according to the specifications [15], the elaboration of the "nonce" is the following one:

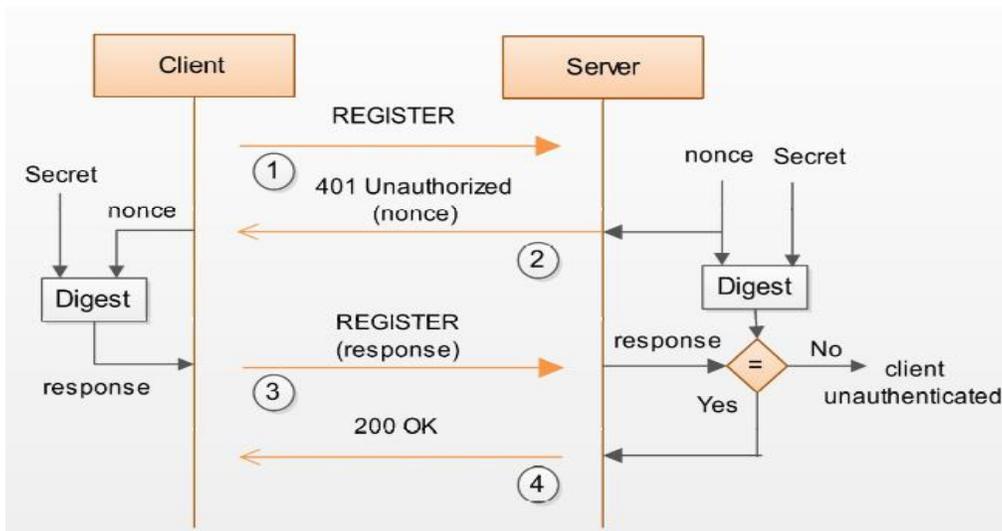


Fig. 3. Classical Registration in the IMS Network.

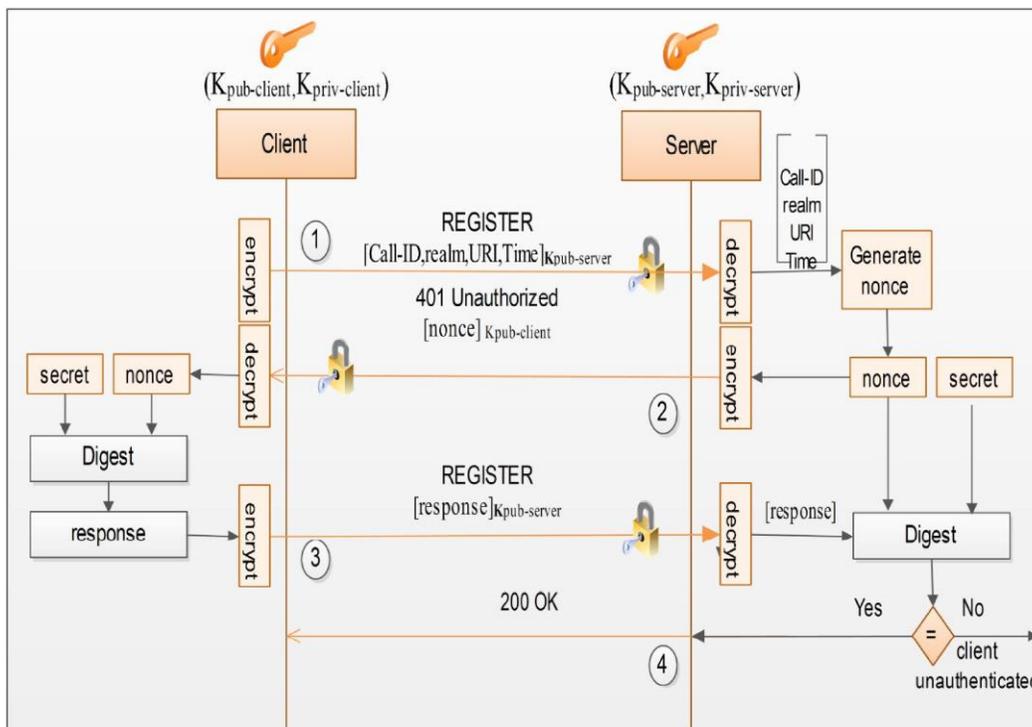


Fig. 4. Enhanced Registration in the IMS.

The registration process offers several advantages, the most important is the mutual authentication, to ensure the integrity and confidentiality of the information exchanged during the registration phase between the client and the server. This procedure also enhances security and prevents attacks that can exploit sensitive information: nonce, call-ID, domain and response. Note that the communication architecture and messages remain as they were before [16].

C. Modelling the New Approach

The specialization of a language (SDL) follows a three-step methodology: specification, design and implementation with code generation. The first step concerns the expression of

constraints and needs by the specification. At this level several languages can be used as UML, SDL. The second step concerns the definition of the execution model it is the design stage. At this level, languages like SDL-RT, LACATRE, UML-RT can be used.

The methodology must follow four essential steps:

- a) Definition of constraints and specifications
- b) Definition of the Structural model
- c) Definition of the Behavioral Model
- d) Verification and Validation

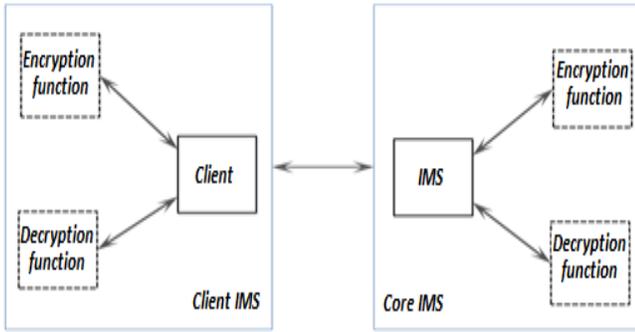


Fig. 5. All Blocks and Sub-Blocks of System.

The objective is to first enhance security by validating the registration process approach of the REGISTER SIP method.

The specifications of SDL must take into consideration all the compositional structures and links in the NGN and especially the IMS, it allows to give a concrete view in the implementation of an external operation launched by a customer such as registration or service request. In this formalism, two resulting processes: a block for the IMS network with all these components, and the second block is the client with all its activities. Likewise, each block is subdivided into three blocks relating to the authentication operation processed. The IMS block contains a master block with the functions (CSCF) of the IMS with two sub-blocks for the encryption and decryption of sensitive information. The client block has a sub-block to simulate the functionality of a client in terms of registration, the other two block aims to encrypt / decrypt information. The blocks in Fig. 5 and the sub-blocks communicate with each other via interfaces and messages [11].

The structural model is a static view, which represents the relationship between modelled entities, their interfaces, and attributes according to SDL. The communication channels between the different block instances specify the signals as SIP messages between the clients and the IMS core. Blocks and processes are used to represent entity types such as client and core IMS with the Cinderella tool. The two main IMS entities according to the definition of constraints and specifications are shown in Fig. 6.

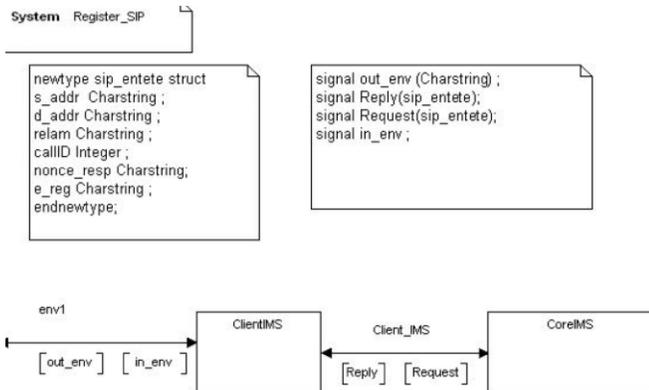


Fig. 6. System Interaction between ClientIMS and CoreIMS.

The model contains the elements:

- **IMS Client Block**
- **Block Core IMS**
- **4 signals:** outenv, inenv, reply and request
- **2 interfaces env1 and clientIMS:** to ensure communication between the entities and their environment.

The system contains mainly four sub-blocks, these blocks are detailed according to their communication and their role of each sub-block:

- **Client Block:** This system block represents all IMS clients, these clients communicate with their environment by signals that activate processes within the block. To register in the IMS network, the block sends a signal to the Client block to activate the process. After a phase of external exchanges, the client process informs the environment of the result received from the IMS core. The process inside the client block contains the different communication interfaces and the signals between this process and the external environment.
- **Block CoreIMS:** The block is very important for the IMS, it includes all the functionalities of the entities: P-CSCF, I-CSCF, S-CSCF and HSS. The signals and the behaviors of each entity are taken into consideration in order to give a better modeling of the block. The block contains only a single process inside it is the process "IMS_process" which ensures the task of recording as well as the interactions between the 4 IMS core entities. The "CI_pr" interface ensures communication between the process and the external environment (Client or Application Server).
- **Block encryption function:** Asymmetric encryption is a communication protocol based on asymmetric mathematical functions with private keys. These mechanisms make it possible to obtain functionalities: confidential data protection, a digital signature or the exchange of secrets [16]. In an RSA crypto-system, each IMS client element and IMS core must build its own RSA module like [17]:

Algorithm 1: Module Manufacturing
<p>Input : A size t for the RSA cryptosystem module. Output : An RSA N module of size t.</p> <ol style="list-style-type: none"> 1: Take a first random number p in the range $[2^{\frac{t}{2}}, 2^{\frac{t+1}{2}}]$, 2: Take a random prime number q in the meantime $[2^{\frac{t}{2}}, 2^{\frac{t+1}{2}}]$, 3: If p = q 4: Next step 2. 5: Else if 6: N = pq. 7: Fi

p and q are in the meantime $[2^{\frac{t}{2}}, 2^{\frac{t+1}{2}}]$, so we have $2^t < pq < 2^{t+1}$

Which shows that N = pq is size t.

After having made an RSA module, each of the participants must prepare a secret key d and a public key e :

In some cases, specific values may be taken for the public key for example: $e = 3$ ou $e = 2^{16} + 1$. In this case, steps 2 to 5 of algorithm 2 are not executed.

The function ϕ plays a central role in the RSA cryptosystem and is called the Euler function. Definition of the function of Euler: Let n be an integer. The indicator function (2) of Euler is:

$$\phi(n) = \#\{a | 0 \leq a \leq n - 1, \text{pgcd}(a, n) = 1\} \quad (2)$$

<p>Algorithm 2: Key Making</p> <p>Input : Two prime numbers p and q. Output : A private key d and a public key e</p> <p>1: Calculate $\phi(n) = (p - 1)(q - 1)$. 2: Prendre un nombre aléatoire e dans l'intervalle $[1; \phi(N)]$. 3: If $\text{pgcd}(e, \phi(N)) \neq 1$ 4: Next step 2. 5: Ifelse 6: Calculate $d \equiv e^{-1}(\text{mod } \phi(N))$ 7: Fi</p>

The function is defined for any integer $n \geq 2$. If the decomposition into primitive factors is

$$n = \prod_{i=1}^s p_i^{x_i}$$

So we have $\phi(n) = \prod_{i=1}^s p_i^{x_i} (p_i - 1)$

<p>Algorithm 3: Encrypting a message</p> <p>Entry: a clear message and the public key. (N_B, e_B). Output: An encrypted message C</p> <p>1: Transform the message into an integer M of the interval $[2, N_B]$ 2: Calculate $C \equiv M^{e_B}(\text{mod } N_B)$. 3: Send the message C</p>
--

- **Block decryption function:** If the IMS receives an encrypted message C from the client. Then the decryption in message B is done using its secret key d_B as in this algorithm:

<p>Algorithm 4: Decrypting a message</p> <p>Entry: an encrypted message C and the private key (N_B, d_B). Output: A clear message M.</p> <p>1: Calculate $M \equiv C^{d_B}(\text{mod } N_B)$ 2: Transform the number M into a clear message.</p>
--

III. IMPLEMENTATION

A. Description of Testbed

The test bench contains four layers of NGN and implements open source solutions (Fig. 7).

1) *Service layer:* The layer contains control entities to access IMS services (CSCF) with open source OpenIMSCore [18]; this layer can expose two types of service, the IMS service with VoD AS [19] - UCT IP TV the IMS server.

Video on Demand (VoD) services allow users to watch video content of their choice at a time. And another traditional Web server Iperf [20], Iperf and an open source performance measurement tool used to test the bandwidth, the bit rate between two hosts so that one host acts as a server and the other as a client. Performance parameters can be measured with either TCP or UDP packets.

2) *Control layer:* This layer exposes control services with the HSS database, as well as the two QoS political entities (PCRF, PCEF) implemented in java code called UCT Policy Control Framework PCRF [21]. The layer also contains a controller that allows QoS management in SDN. We choose Floodlight [22] is a range of the Beacon Controller, a Java-based OpenFlow Controller with Apache license. FloodLight is chosen as the OpenFlow controller to be used to coordinate stream inputs and the NGN / IMS architecture. This controller is chosen because it is a robust and powerful controller.

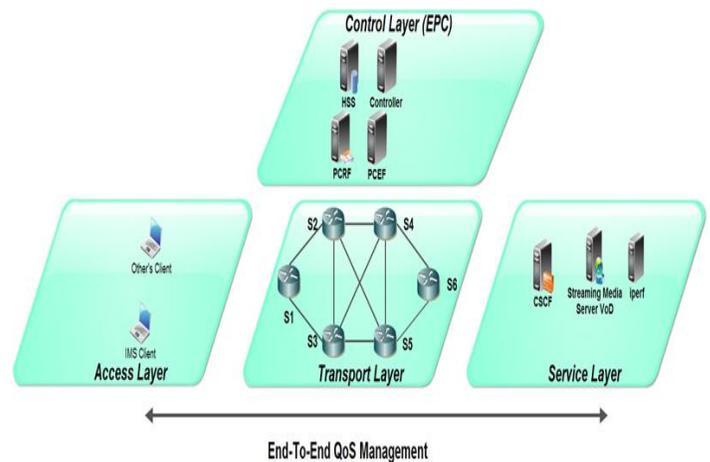


Fig. 7. TestBed Architecture.

3) *Transport layer:* This layer contains a set of switches that provide the interconnection with the two access and service layers. The topology is created with Mininet [23]. Mininet is used in this project to emulate the topology and to test traffic flows. A Python script is used to create the topology in Mininet, and traffic streams are received from a remote OpenFlow controller. In this layer, we defined two types of core and edge switches to implement DiffServ. The edge switch provides classification, measurement, queuing and scheduling operations; The core switch, at the input, performs a class flow and applies the PerHop Behavior (PHB). The TC traffic control tool [24] is used to show and manipulate traffic control settings. The queues are deployed in TC to ensure that each queue receives the level of service required for its class. A set of queue disciplines is implemented: First-In-First-Out (FIFO), pending class (CBQ) [24], HTB [25]. And iptables to classify all packages.

4) *Access layer:* The layer contains ordinary clients and an IMS client, the latter is UCT IMS Client [26] - The IMS client. The UCT IMS client support VoD / IPTV services. Our objective in this test bed is to secure client WiFi access to

server via IMS network. For that we have to perform two actions:

- The Wi-Fi Client/User authentication: we use a centralized authentication server RADIUS [27] with EAP/TLS.
- Secure Client/Server communication: basically, SIP and RTP flows.

The Confidentiality, the integrity and the mutual authentication are the services we need to achieve our goal. We chose to use IPsec tunnel because it's the best advantage of securing all applications data and media transparently in IP layer. The test bed implements IPsec on tunnel mode with ESP as security protocol, AES-128 as algorithm for confidentiality, SHA-1 as algorithm for integrity, and pre-shared key for mutual authentication. And also, our solution proposed is integrated into OpenIMScore and Clients.

B. Experience & Test

The objective of the test is to verify the impact of the SIP enhancement solution in the authentication operation, although check the correct integration in the IMS opensource solution. According to the test the authentication is done correctly with our new method, it remains to check the direct impact of the authentication mechanism on the response time as key of QoS.

For this it was necessary to highlight the platform response for a consistent number of users with several security solutions proposed which gives Fig. 8. The graph represents the response time or registration delay is represented in tree case: None, IPsec, TLS, and finally with SIP_Enhanced which is represent our proposition for enforcing security in SIP.

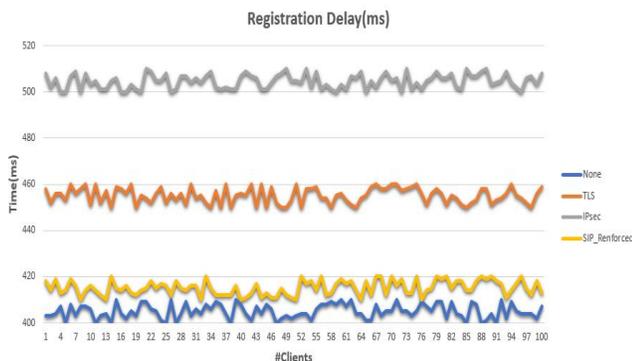


Fig. 8. Registration Delay in IMS: None, TLS, IPsec and SIP_Enhanced.

The measured time is End-to-End between the client and the platform. The solutions proposed with TLS and IPsec contain security problems as well as a significant delay relative to the SIP protocol. On the other hand, the delay is less important in SIP authentication, especially with our solution that is close to values without the use of a security mechanism.

IV. CONCLUSION

Registration with an NGN is a sensitive and necessary operation before the request for any services. The new 4G or 5G generation offers SIP, because of its simplicity, as a primary signalling protocol. Nevertheless, in the basic registration operation, the same vulnerabilities exist with

security solutions like IPsec or TLS for SIPs, these mechanisms have an impact not only by the lack of total security but also on the QoS and especially the time of reply.

In our work, instead of using traditional solutions like IPsec and TLS, we do an analysis of the NGN architecture as well as the registration procedure with SIP. The study shows that it is possible to modify the key supply principle (private and public) without changing the exchange messages or the protocol followed by authentication. After the analysis, comes the modelling step that we used SDL to integrate our solution with test bed to first test the feasibility and actual operation of our proposal and second to know the impact of direct on the response time.

The integration is done without problem as well as the measurement results show that the delay is close to what existed before. Our next research jobs focus on other type of signalling and other SIP message or other usual service in NGN (5G). The security aspect is still a persistent problem in the IP world.

REFERENCES

- [1] https://www.itu.int/ITU-T/studygroups/com13/ngn2004/working_definition.html
- [2] B.Raouyane, M.Bellafkih, D.Ranc, "QoS management in IMS: DiffServ model". NGMAST 2009: 3rd International Conference and Exhibition on Next Generation Mobile Applications, Services and Technologies, IEEE Computer Society, 15-18 september 2009, Cardiff, Wales, United Kingdom, 2009, pp. 39-43, ISBN 978-0-7695-3786-3.
- [3] B.Raouyane, M.Bellafkih, M.Errais, M.Ramdani, "IMS management and monitoring with eTOM framework and composite web service", International Journal of Next-Generation Computing (IJNGC) - ISSN 2229-4678, eISSN 0976-5034 Vol. 2, No. 2, 2011.
- [4] ETSI TS 123 272 V13.3.0 (2016-04) Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2 (3GPP TS 23.272 version 13.3.0 Release 13
- [5] SDN Architecture Overview Version 1.0 December 12, 2013 available via <https://www.opennetworking.org/wp-content/uploads/2013/02/SDN-architecture-overview-1.0.pdf>
- [6] OpenFlow Switch Specification Version 1.5.1 (Protocol version 0x06) March 26, 2015 ONF TS-025, available via <https://www.opennetworking.org/wp-content/uploads/2014/10/openflow-switch-v1.5.1.pdf>
- [7] Bellafkih.M. Belmekki.E. "efficient light model for securing IMS network". Intelligent Systems: Theories and Applications (SITA), 9th International Conference on Date 8-9 May 2013. EMI. Rabat Maroc IEEE, (8-9), May 2013.
- [8] Bouaouda.N; Raouyane.B; Belmekki.E; Bellafkih.M. "IP Multimedia Subsystem : Security evaluation". Journal of Theoretical and Applied Information Technology, Vol. 51, No.1(10), May 2013.
- [9] Belmekki.E. " Analyse des risques par EBIOS et validation d'une approche pour la sécurisation dans un réseau IMS". thèse de doctorat soutenu a la FST Mohammedia, Université Hassan 2, septembre 2015.
- [10] 3GPP. "network domain security, IP network layer security". TS 33.210 (Release 5), March 2008.
- [11] Garcia-Martin.M. Camarillo.G. "the 3g IP Multimedia Subsystem (IMS) : Merging the internet and the cellular worlds". John Wiley et Sons, ISBN : 9780470516621, 2008.
- [12] ITU-T. "recommendation z.100. specification and description language (sdI).technical report z-100". International Telecommunication Union Standardization Sector, Genève, 1994.
- [13] Bellafkih.M. Belmekki.E, Raouyane.B. "secure sip signalling service in IMS network". Intelligent Systems: Theories and Applications (SITA),

- 9th International Conference on Date 8-9 May 2014.Rabat Morocco IEEE, (8-9), May 2014.
- [14] Bouaouda.N; Belmekki.E; Bellafkih.M. "towards a new approach for securing IMS networks". AASRI Conference on Intelligent Systems and Control (ISC 2013) Vancouver, Canada, proceeding published by AASRI Procedia (ISSN : 2212-6716) by ELSEVIER, which will be indexed by ScienceDirect and Scopus, Vol. 4(17-18), April 2013.
- [15] J. Rosenberg And el , "SIP: Session Initiation Protocol", RFC 2361, June 2002
- [16] Belmekki E.; Bellafkih M.; Belmekki A. «Enhances security for IMS client» Fifth International Conference on Next Generation Networks and Services (NGNS) 28-30 May 2014, Casablanca, Morocco IEEE
- [17] Dobbartin.H."The status of md5 after a recent attack". cryptoBytes, vol 2 n.2(p 1-6), 1996.
- [18] OpenIMScore home page, [online] Available: <http://http://www.openimscore.org/>
- [19] Uct ip tv home page, [online] Available:<https://linuxstgo.wordpress.com/2012/04/30/how-to-setup-uct-advanced-iptv/>
- [20] IPerf home page, [online] Available: <https://iperf.fr/>.
- [21] PCRF home page, [online] Available: https://developer.berlios.de/project/showfiles.php?group_id=7844
- [22] Floodlight Queues home page, [online] Available: <https://floodlight.atlassian.net/wiki/display/floodlightcontroller/How+to+Use+OpenFlow+Queues>
- [23] Mininet home page, [online] Available: <http://mininet.org/walkthrough/>.
- [24] Linux TC home page, [online] Available: <http://linux.die.net/man/8/tc>.
- [25] Linux HTB home page, [online] Available: man7.org/linux/man-pages/man8/tc-cbq.8.html
- [26] D. Waiting, R. Good, N. Ventura, "The UCT IMS Client", 2008. Uct ims client home page, [online] Available: <https://yulexs.wordpress.com/tag/uct/>.
- [27] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.

Information System Quality: Managers Perspective

Sarah Aouhassi, Mostafa Hanoune

Laboratory of Information Technology and modeling,
Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca
B.P. 7955 Sidi Othmane, Casablanca, Morocco

Abstract—To evaluate Information System Quality (ISQ) quantitatively, a model was constructed based on sub-models related to the five Information System (IS) components, namely, Human Resources, Hardware, Software and application, Procedure and Data, and all IS players perspectives are considered who are: Managers, Technical Staff, Functional Staff and Users. This paper focuses on the survey designed for managers in order to form the variable indicators from variable questions, via appropriate formulas in the first place, and to analyze data collected from IS managers of the Moroccan universities in the second one. This approach will allow diagnosing precisely the malfunctioning areas on ISQ by emphasizing on the components with less quality level. It will also enable making comparison of ISQ on different organizations with the mean of standardized values.

Keywords—Information system; quality; managers; measurement indicator; university

I. INTRODUCTION

The first thing one think about when approaching ISQ field is software quality with all its inherent models and practices, characteristics and sub-characteristics, factors and criteria, metric and measure,[5; 6; 9; 12; 13; 14; 15; 19]. However, IS is defined as an organized set of resources (human, software, hardware, procedures and data) which allow to collect, sort, classify, treat and transmit information on working environment, therefore, IS quality should be a reflection of the quality level of all its components [22].

The literature review on the field of ISQ shows a variety of models that treat this problem; each one is focusing on a number of features with a multitude of metrics. However a common limitation rises, firstly, all models mix up ISQ with Software and Application Quality (SAQ), secondly, the surveys used to collect data are designed basically for technical staff only.

On previous work [1], a global model was defined and named ISysQ with a set of measuring indicators covering all IS attributes for each component (Fig. 1) and customized surveys were constructed and adapted to the respective IS intervening (IS managers, technical staff, functional staff and users). The global model contains 25 indicators as mentioned in Fig. 1, but not all these indicators concern at the same time every IS player.

The focus of this study is the managers' survey [20] and their perspectives about ISQ [11]. The IS managers are the party who is meant to ensure quality of IS in any organization. In fact, manager is by definition "An individual who is in charge of a certain group of tasks, or a certain subset of a

company. A manager often has a staff of people who report to him or her. Certain departments within a company designate their managers to be line managers, while others are known as staff managers, depending upon the function of the department. (<http://www.businessdictionary.com>). The definition of quality by ISO 8402-1986 standard is "the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs". Crossing these definitions with the one of IS, it will be implied that IS managers are the players who are supposed to have inclusive information about ISQ.

This study aims to give numerical values instead of qualitative description to IS indicators and components [2]. The numerical values attributed to variable questions that constitute the survey are aggregated in variable indicators [16] by component, giving rise to summary values of ISQ components. This approach allows to highlight strengths and weaknesses of each IS component in order to provide later corrective measures. ISQ quantification will allow thereby an objective comparison of several organizations [7] in detailed way by going down to the lowest level of the hierarchical structure of ISysQ which are indicators.

The ISysQ model have five dimensions related to the five IS components. In this contribution, a particular focus will be on the Human Resources Quality and Data Quality dimensions. The same approach can be applied to the other ones.

In the following sections, first the theoretical model of ISQ with 25 indicators is developed. Then, the specification of the model adapted for IS Managers is performed. After that, a presentation of the research methodology including the description and implementation of the study on one hand and the aggregation of the variables questions into variable indicators on the other hand is realized. Next, the analysis and results of the data collected is presented. Finally, the paper concludes with a discussion of the findings and directions for future research.

II. THEORETICAL BACKGROUND AND CONTEXT OF IS QUALIFICATION

The literature review in the field of IS quality has shown a major deficiency related to the IS components other than software and applications [10; 19], while having an IS quality means that all its components have a certain quality level.

The hybrid model adopted [1] is composed of five sub-models as shown in Fig. 1.

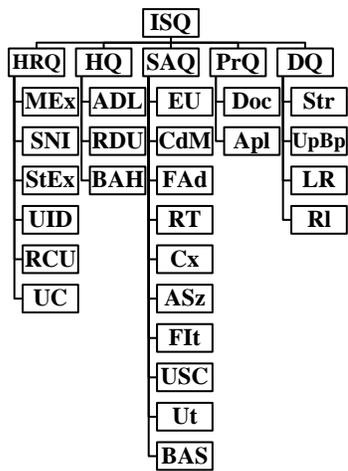


Fig. 1. IS Quality Model Indicators.

Among the 25 indicators that constitute the ISQ model, 22 are related to the IS Managers as mentioned in Fig. 2, reaching thereby the maximum of indicators that can have an IS intervening sub-model.

The indicators (gathered by IS component) which don't concern IS Managers are:

- **HRQ:** User competency (UC).
- **HQ:** Average duration of life (ADL)
- **SAQ:** Complexity (Cx)

The reason why these indicators are excluded from the IS Managers sub model is that they can't answer corresponding questions. Information that is purely technical e.g. ADL and Cx or relative to Users e.g. UC, is to be eliminated from the IS Managers survey [8]. On the other hand, IS Managers give relevant and precise information about their expertise areas such as budget allocated to hardware, documentation quality or details about the staff involved in IS.

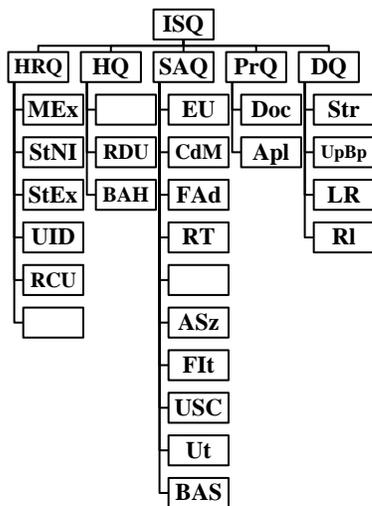


Fig. 2. Indicators Relative to IS Managers.

III. RESEARCH METHODOLOGY

A. Study Description

The surveys are designed in order to be adapted for each type of the questioned: IS Managers, technical staff, functional staff and users. The survey first part, regardless of type, helps to make a profile picture of the respondent, the second part deal with IS generalities, e.g. the IS department size, in numbers and staff skills or qualification. The third part emphasizes the relationship between the respondent and other IS contributors, like the difficulty met when detailing technical requirements by managers for developers. The last part of the survey is about measuring indicators concerning software/application and hardware utilization in order to see if there is a way to optimize available resources, beside software and application impact on reduction time on performing a given task and on IS contributors' efficiency. The structure above is common to the four types of surveys; nevertheless every survey has its own distinctive feature specific to the different kind of staff, subject of the inquiry.

The survey addressed to IS manager focuses on the governance side of information system like the allocated budget for IS structure, the global strategies or orientations of the firm.

B. Study Implementation

The case study is Moroccan public universities where the surveys were distributed to the IS managers and a group of technical staff during a national meeting organized within the project TEMPUS¹ MISSION² on October 29th and 30th 2015 in Agadir Morocco. In such a context, the respondents are naturally engaged and the collected data are reliable. Data from other IS intervening were collected from each university via the IS manager and sent by email. The majority of questions contained on surveys return quantitative data, except a few questions which allow qualitative answers to help understanding and clarifying some subjects. A primary study and analyze of the surveys was presented during the closing meeting of the project on March 11th 2016 (<https://goo.gl/UArq4t>), and data complement was collected just shortly after by email. In the following, the aggregation of the survey questions into the model indicators will be pointed out and followed up with data analysis.

C. Aggregation of Questions Into Indicators

1) Human resources quality (HRQ)

a) Manager Experience (MEx)

The IS quality is directly affected by the IS manager experience [17]. Decisions and strategies adopted are determining for the whole IS intervening. This indicator is measured in term of years of experience on management and IS and aggregates the questions below:

- How many years' experience on Management do you have? (ExM)

¹ Trans-European Mobility Program for University Studies

² Establishment of a National Service of an Operational Information System

- How many years' experience on IS do you have? (ExSiM)

The variable indicator is a mean of the two variables questions as mentioned on the formula below:

$$ME_x = \frac{(ExM + ExSiM)}{2}$$

The levels of each variable are described in Table II.

b) Staff numbers involved in IS (StNI)

This indicator returns the number of the staff involved in IS including every one that contributes directly or indirectly on IS development, categorized by profiles, which are competence degree and IT specialization. It aggregates the questions below:

- What is the number of the following profiles composing the IS department?
 - IT specialist (EfInf)
 - Non IT specialist (EfNInf)
 - Executives (EfCdr)
 - Middle Executives (EfCdrm)
 - Technician (EfTch)

The indicator value is the mean of the weighted variables questions as indicated on the following formula:

$$StNI = \frac{(2 \times EfInf + EfNInf) + (3 \times EfCdr + 2 \times EfCdrm + EfTch)}{2}$$

The variable questions above are subdivided on two groups, the first one divides the IS staff in two categories (IT specialist and Non IT specialist) and the second one divides the IS staff in three categories (Executives, Middle Executives and Technician). The weight of each variable question is equal to its importance degree on the group.

c) IS staff experience (StEx)

The experience accumulation of IS staff lead to a better quality of IS itself through avoiding frequent errors and reducing task's length. This indicator is measured in term of years of experience and competence degree of IS staff. It aggregates the questions below:

- What is the number of the following profiles composing the IS department?
 - Staff with less than 2 years' experience (Ef2a)
 - Staff with experience between 2 and 5 years (Ef2a5a)
 - Staff with experience between 5 and 10 years (Ef5a10a)
 - Staff with more than 10 years' experience (Ef10a)
- How do you evaluate the technical staff's skills? (CompTch)
- What is the number of functional staff? (EffFct)

- How many applications are developed by the technical staff? (NbDev)

The formula relating these variable questions to the corresponding variable indicator is as following:

$$StEx = \frac{1}{4} \times ((Ef2a + 2 \times Ef2a5a + 3 \times Ef5a10a + 4 \times Ef10a) + (4 - CompTch) + EffFct + NbDev)$$

Since the indicator is by definition reflecting the experience degree of IS staff, the number of IS staff of each category is weighted by years of experience. The minus sign appears to conserve the logical order of the variables level (Table II). The formula above returns values included between 3 and 10 as detailed in Table I.

TABLE I. RELATING STEx COMPUTED VALUES WITH ANSWERS

Variable Indicator	Interval	Value	Answers
StEx]3,4]	1	Inexperienced
]4,6]	2	Less Experienced
]6,8]	3	Averagely experienced
]8,10]	4	Experienced

d) Users implication degree (UID)

This indicator is measured by the number of interactions with available applications and software [21; 23].

- Are there any unused applications? (ApNUt)
- If yes, how many unused applications are there? (NbApNUt)

$$UID = \begin{cases} NbApNUt & \text{if } ApNUt = 1 \\ 4 & \text{if } ApNUt = 2 \end{cases}$$

e) Resistance to change of users (RCU)

RCU measures the adherence degree of users facing the new practices related to IS [23]. This indicator is expressed on the survey by the question below and takes the same values as those of Ad (Table II).

- What is the adherence degree of users to new information system practices? (Ad)

2) *Hardware quality (HQ)*

a) Rate of daily use

The number of hours past at using IT equipment divided by the number of daily work hours. This indicator is sorted by hardware type (computer, printer, server...). The corresponding questions are as below:

- What is the average number of hours per day spent on using hardware type i? (RDU)

$$i \in \{\text{computer, printer, server, scanner, ...}\}$$

b) Budget allocated to hardware (BAH)

This indicator gives an indication of the budget allocated to hardware using the question below:

- What is the portion of the budget allocated for hardware's purchase and maintenance? (BgAM)

The proportion is used instead of the real amount to allow later comparison between organizations.

3) Software and application quality (SAQ)

a) Ease of use (EoU)

The exploitation rate of software and applications gives a numerical indication for the ease of use noted by the different players. For instance, as managers have an overview of all the software and applications available on the organization, they can answer this question thoroughly.

- What is the exploitation rate of the existing software and applications? (TExp)

b) The code development maintainability (CDM)

Maintainability of the code development allows saving time and energy, and thereby contributes on improving the IS quality [18]. The questions corresponding to this indicator are:

- Has the code been reused for other applications? (CdRut)
- If yes, specify the original application and the destination one! (ApOr1, ApDst1, ApOr2, ApDst2, ApOr3, ApDst3).

(The survey allows three possibilities for the question above).

The formula relating these variable questions to the variable indicator is as following:

$$CDM = \begin{cases} 0 & \text{if } CdRut = 2 \\ \sum_{1}^3 ApOri & \text{if } CdRut = 1 \end{cases}$$

The value of the indicator CDM is set to null when the answer is that the code is not reused for other applications and it takes the sum of reused code application if the answer is yes (Table III).

c) Flexibility or adaptability (FAd)

The ability of software and applications to satisfy similar needs to requirements originally specified. This indicator is reported on the survey by the question below:

- Do you think that available software and applications can meet similar needs to those initially specified? (FAd)

The values of this variable indicator are the same as those of the related variable question (Table III).

d) Response time (RT)

The duration between the time the request is executed and the response time, this indicator is reported on the survey by the question below:

- How do you assess the software and applications response time?

This indicator is measured qualitatively and its values vary from very slow to very fast (Table III).

e) The application/software size (ASz)

The size of an application can be measured in different ways, but the most appropriate way to find out from a manager is the total time spent on programming, formulated on the survey by the question below:

- What is the total programming time for an application? (DurT)

f) Friendly interfaces (Fit)

The interfaces should be practical and intuitive according to user's opinion. This indicator is reported on the survey by the questions below:

- Are the software/application interfaces friendly? (Fit)
- If no, explain why! (FitN)

The indicator takes the value of the first question and uses the answer of the second one as a clarification.

g) Users specifications conformity (USC)

Developed applications or software have to match with the requirements initially specified, this indicator is reported on the survey by the question below:

- Are the developed applications compliant with the original specifications? (USC)

The indicator takes the values: yes/ partly /no.

h) Utility (Ut)

The gap between the situations with and without the software, in terms of efficiency and work duration. This indicator is staggered from 1 "no utility" to 5 "very useful". The questions related to this indicator are:

- How useful is the application / software in terms of working time? (UtTp)
- If the application / software have not induced any change in working time, explain why! (UtTpN)
- How useful is the application / software in terms of efficiency? (UtEf)
- If the application / software have not induced any change in efficiency, explain why! (UtEfN)

The formula aggregating the variables question into variable indicator is as below:

$$Ut = \frac{(UtTp + UtEf)}{2}$$

It is to be noted that the variables question UtTpN and UtEfN are qualitative and their roles is limited to enlighten why the introduction of software and applications didn't produce any positive impact in terms of efficiency and work duration.

i) Budget allocated to software and application (BAS)

The proportion of the annual budget spent on new software and/or on application development.

- How much software were acquired? (NbLog)

- What is the portion of the budget allocated for software’s purchase? (BgLog)
- What is the portion of the budget allocated for staff training involved on the Information system? (BgPSI)

TABLE II. RELATING QUESTION VARIABLES ON MANAGERS SURVEY TO HUMAN RESOURCES QUALITY INDICATOR VARIABLES

Variables indicator	Answers	Values	Variables question	Answers	Values
MEX	Inexperienced	1	ExM	Less than 2 years	1
	Less Experienced	2		Between 2 and 5 years	2
	Averagely experienced	3		Between 5 and 10 years	3
	Experienced	4		More than 10 years	4
StNI	Small number	1	EfInf	Less than 2 years	1
				Between 2 and 5 years	2
				More than 10 years	3
	Average number	2	EfNInf	Less than 5 persons	1
				Between 5 and 10 persons	2
				More than 10 persons	3
	Sufficient number	3	EfCdr	Less than 5 persons	1
				Between 5 and 10 persons	2
				More than 10 persons	3
			EfCdrm	Less than 5 persons	1
				Between 5 and 10 persons	2
				More than 10 persons	3
StEx	Inexperienced	1	Ef2a	Less than 5 persons	1
				Between 5 and 10 persons	2
				More than 10 persons	3
	Less Experienced	2	Ef2a5a	Less than 5 persons	1
				Between 5 and 10 persons	2
				More than 10 persons	3
	Averagely experienced	3	Ef5a10a	Less than 5 persons	1
				Between 5 and 10 persons	2
				More than 10 persons	3
	Experienced	4	CompTch	High skills	1
				Average skills	2
				Low skills	3
UID	No implication	1	EffFct	Less than 5 persons	1
	Low implication	2		Between 5 and 10 persons	2
	Average implication	3		More than 10 persons	3
	High implication	4		None	1
RCU	No adherence	1	NbDev	Between 1 and 5 app	2
	Low adherence	2		Between 5 and 10 app	3
	Average adherence	3		More than 10 app	4
	High adherence	4		Yes	1
UID	No implication	1	ApNUt	No	2
	Low implication	2		More than 10	1
	Average implication	3		[5, 10[2
	High implication	4		[1, 5[3
RCU	No adherence	1	Ad	No adherence	1
	Low adherence	2		Low adherence	2
	Average adherence	3		Average adherence	3
	High adherence	4		High adherence	4

The variable indicator is a mean of the three variables question above and its values are reported in Table III.

$$BAS = \frac{(NbLog + BgLog + BgPSI)}{3}$$

4) Procedures quality (PrQ)

a) Documentation (Doc)

Documentation quality on the literature review refers to the documentation accompanying a software development [3], whereas the indicator Doc in our model reflects one side of procedures quality.

The indicator Doc is reported on the survey by the four questions below:

- Does your organization have an information system blueprint? (SDSI)

- Are there procedures for information system in your organization?(PrSI)
- Is there a specific documentation for IS procedures? (DocSI)
- Is there an entity responsible for production, updating, archiving, etc. of this documentation? (EnDoc)

$$\text{Doc} = \frac{(\text{SDSI} + \text{PrSI} + \text{DocSI} + \text{EnDoc})}{4}$$

This indicator is staggered from “compliant” when all the quality attributes exist to “non-existent” where they are all absent (Table III).

TABLE III. RELATING QUESTION VARIABLES ON MANAGER SURVEY TO SOFTWARE/APPLICATION, PROCEDURE AND DATA QUALITY INDICATOR VARIABLES

Variables indicator	Answers	Values	Variables question	Answers	Values
EoU	Very difficult to use	0	TExp	[0%, 25%]	1
	Difficult to use	1		[25%, 50%]	2
	Moderately difficult	2		[50%, 75%]	3
	Moderately easy	3		[75%, 100%]	4
	Easy to use	4			
	too easy to use	5			
CDM	Not maintainable	0	CdRut	Yes	1
	Maintainable 1	1		No	2
	Maintainable 2	2	ApOri	Blank	0
	Maintainable 3	3		Filled	1
FAd	Yes	1	FAd	Yes	1
	No	2		No	2
RT	Very slow	1	RT	Very slow	1
	Slow	2		Slow	2
	Average	3		Average	3
	Fast	4		Fast	4
	Very fast	5		Very fast	5
ASz	Small	1	DurT	Small	1
	Medium	2		Medium	2
	Large	3		Large	3
	Very large	4		Very large	4
Fit	Yes	1	Fit	Yes	1
	No	2		No	2
Ut	No utility	1	UtTp	Qualitative	
				No utility	1
				Low utility	2
				Average utility	3
				Useful	4
	Very Useful	5			
	Average utility	3	UtTpN	Qualitative	
				No utility	1
				Low utility	2
				Average utility	3
Useful				4	
Very Usful	5				
Useful	4	UtEf	Qualitative		
			No utility	1	
			Low utility	2	
			Average utility	3	
			Useful	4	
Very Usful	5				
BAS	Insufficient	1	NbLog	[0, 1]	1
				[1, 4]	2
				More than 5	3
	Moderate	2	BgLog	Less than 0,1%	1
]0,1%, 10%[2
				More than 10%	3
Sufficient	3	BgPSI	Less than 0,1%	1	
]0,1%, 10%[2	
			More than 10%	3	
Doc	Compliant	1	SDSI	Yes	1

	Non-compliant	2	PrSI	No	2
				Yes	1
			Non existent	3	DocSI
	Yes	1			
	EnDoc				Yes
			No	2	
Apl	Applicable	1	AplPr	Applicable	1
				Partly applicable	2
				Not Applicable	3
	Partly applicable	2	PrRap	Yes	1
				Partly	2
				No	3
RI	No relevance	1	Ind	0	1
				[1, 5[2
				[5, 10[,	3
	Low relevance	2	Res	More than 10	4
				0	1
				[1, 5[2
Average relevance	3		[5, 10[,	3	
			More than 10	4	

b) Applicability (Apl)

The quality of the procedures depends on their applicability by the IS intervening. This indicator is staggered from 1: “applicable” to 3: “not applicable” and aggregates the two questions below:

- What is the applicability degree of the procedures by the Information System intervening? (AplPr)
- Is there any tangible impact of the procedures on the speed of daily tasks? (PrRap)

$$Apl = \frac{(AplPr + PrRap)}{2}$$

5) *Data quality (DQ)*: The quality of data is a “multi-dimensional measure of the suitability of data to fulfill the purpose bound in its acquisition/generation. This suitability may change over time as needs change” [3; 4]. This underlines the subjective requirements for data quality in respective institutions and illustrates a possible dynamic data quality process. The definition makes it clear that “the quality of data depends on the time of the consideration and on the level of claims placed at the time on the data”.

a) Structure(Str)

- Are the data stored in a DBMS?

b) Updating and back up(UpBp)

- What is the time interval between two backups?

c) Lack of redundancy(LR)

- Are there any data redundancies?

d) Relevance (RI)

- What is the number of indicators serving the objectives drawn by the University? (Ind)

- What is the number of expected results from these indicators? (Res)

$$RI = \frac{(Ind + Res)}{2}$$

IV. DATA ANALYSIS AND RESULTS

Data analysis was performed through three phases. First, row data from the IS Managers survey were gathered by indicator. Second, the aggregation formulas were used to give the indicators numerical values. Third, a standardization of all values was performed in order to have summarized value for each component and to compare later between the universities subject of the study. It should be noted that because of length restriction, only HRQ and DQ dimensions will be analyzed on the following.

A. Indicator Quantification

The central objective of the model ISysQ is to have numerical values for every indicators, components and finally for the ISQ as a whole. Table IV gives indicator values computed from data collected via the survey designed for IS managers of Moroccan public universities, and then aggregated by the formulas defined previously.

The range of the indicator Manager Experience (MEx) is from Inexperienced (coded as 1) to Experienced (coded as 4) (Table 5.1.), it’s noted that UAE and UIT have the highest value concerning this indicator.

Focusing on the remaining indicators of HRQ lead to a finding that for each one there is a different university that has the highest level, so it’s not possible nor to compare Universities according to all HRQ indicators simultaneously and determine the university with the highest level of HRQ, neither to aggregate the indicator values on one value for HRQ because of indicators range difference, thus the necessity to have comparable values for all the indicators.

TABLE IV. INDICATOR VALUE BY UNIVERSITY

University	HRQ					DQ			
	MEx	StNI	StEx	UID	RCU	Str	UpBp	LR	RI
USMBA	3,5	8	5	3	3,67	2	3	2	3,67
UMP	3,5	9	4,75	3	2,67	2	3	2	2,17
UIZ	1	5,5	4,5	3	3,67	2	3	2	1
USMS	3,5	4,5	3,75	3	2,67	1	1	1	2,17
UAE	4	7	5	3	2	2	1	2	2
UIT	4	9	5,25	4	3,33	2	1	2	3,33
UMI	2,5	5,5	4,5	3	3	2	1	1	1
UCD	3,5	7	5	4	2,67	2	3	2	2,5
UHI	3,5	7	5,5	3	4	2	3	2	3
UMV	3	8,5	4,75	3	3	2	1	1	2,33
UHIIC	2,5	8	5,25	3	3	2	1	1	2
UCA	2,5	6	5	4	3,33	2	2	2	3

TABLE V. TABLE 1.1. INDICATOR LEVEL OF HRQ SUB MODEL BY UNIVERSITY

University	MEx	StNI	StEx	UID	RCU
USMBA	Averagely experienced	Average number	Less Experienced	Average implication	Average adherence
UMP	Averagely experienced	Average number	Less Experienced	Average implication	Low adherence
UIZ	Inexperienced	Average number	Less Experienced	Average implication	Average adherence
USMS	Averagely experienced	Small number	Inexperienced	Average implication	Low adherence
UAE	Experienced	Average number	Less Experienced	Average implication	Low adherence
UIT	Experienced	Average number	Less Experienced	High implication	Average adherence
UMI	Less Experienced	Average number	Less Experienced	Average implication	Average adherence
UCD	Averagely experienced	Average number	Less Experienced	High implication	Low adherence
UHI	Averagely experienced	Average number	Less Experienced	Average implication	High adherence
UMV	Averagely experienced	Average number	Less Experienced	Average implication	Average adherence
UHIIC	Less Experienced	Average number	Less Experienced	Average implication	Average adherence
UCA	Less Experienced	Average number	Less Experienced	High implication	Average adherence

TABLE 5.2. INDICATOR LEVEL OF DQ SUB MODEL BY UNIVERSITY

University	Str	UpBp	LR	RI
USMBA	Structured	short	checked	High relevance
UMP	Structured	short	checked	Average relevance
UIZ	Structured	short	checked	No relevance
USMS	Not Structured	depending	Not checked	Average relevance
UAE	Structured	depending	checked	Low relevance
UIT	Structured	depending	checked	High relevance
UMI	Structured	depending	Not checked	No relevance
UCD	Structured	short	checked	Average relevance
UHI	Structured	short	checked	Average relevance
UMV	Structured	depending	Not checked	Average relevance
UHIIC	Structured	depending	Not checked	Low relevance
UCA	Structured	long	checked	Average relevance

B. Values Standardization

Given the inability to compare indicators with the actual values, standardization is required. Standardization is the process of putting different variables on the same scale. This process allows comparing scores between different types of variables. Typically, to standardize variables, the mean and standard deviation must be computed for a variable. Then, for each observed value of the variable, the mean is subtracted and divided by the standard deviation.

Tables 6.1 and 6.2 gives standardized values for HRQ and DQ indicators according to the method described above, the last column contains aggregated value of the components where the value that takes the component is the mean of standardized values of indicators constituting it.

It can be noticed from data in Tables 6.1 and 6.2 that indicators values become comparable and the component value provide a summarized information about the quality state of the IS component.

TABLE VI. TABLE 2.1. STANDARDIZED INDICATOR VALUE OF HRQ BY UNIVERSITY

University	MEx	StNI	StEx	UID	RCU	HRQ
USMBA	0,51	0,65	0,33	-0,58	1,11	0,41
UMP	0,51	1,36	-0,24	-0,58	-0,78	0,05
UIZ	-2,56	-1,12	-0,81	-0,58	1,11	-0,79
USMS	0,51	-1,83	-2,52	-0,58	-0,78	-1,04
UAE	1,13	-0,06	0,33	-0,58	-2,05	-0,24
UIT	1,13	1,36	0,90	1,73	0,46	1,12
UMI	-0,72	-1,12	-0,81	-0,58	-0,16	-0,68
UCD	0,51	-0,06	0,33	1,73	-0,78	0,35
UHI	0,51	-0,06	1,47	-0,58	1,73	0,62
UMV	-0,10	1,00	-0,24	-0,58	-0,16	-0,01
UHIIC	-0,72	0,65	0,90	-0,58	-0,16	0,02
UCA	-0,72	-0,77	0,33	1,73	0,46	0,21
Mean	3,08	7,08	4,85	3,25	3,08	
SD	0,81	1,41	0,44	0,43	0,53	

TABLE 6.2. STANDARDIZED INDICATOR VALUE OF DQ BY UNIVERSITY

University	Str	UpBp	LR	RI	DQ
USMBA	0,30	1,14	0,71	1,67	0,95
UMP	0,30	1,14	0,71	-0,22	0,48
UIZ	0,30	1,14	0,71	-1,70	0,11
USMS	-3,32	-0,96	-1,41	-0,22	-1,48
UAE	0,30	-0,96	0,71	-0,44	-0,10
UIT	0,30	-0,96	0,71	1,24	0,32
UMI	0,30	-0,96	-1,41	-1,70	-0,94
UCD	0,30	1,14	0,71	0,19	0,58
UHI	0,30	1,14	0,71	0,82	0,74
UMV	0,30	-0,96	-1,41	-0,02	-0,52
UHIIC	0,30	-0,96	-1,41	-0,44	-0,63
UCA	0,30	0,09	0,71	0,82	0,48
Mean	1,92	1,92	1,67	2,35	
SD	0,28	0,95	0,47	0,79	

V. DISCUSSION

A. Comparing Universities According to the IS Component "Human Resources (HR)"

Among the twelve Moroccan public universities, UIT is the best in term of HRQ according to IS managers, with a score of 1,12, while USMS has the lowest HRQ score (-1,04). Let's take the case of two universities that have scores close to the mean: UHIIC (0,02) and UMV (-0,01), and try to pursue steps to an inclusive comparison. First of all, when taking the qualitative attributes of HRQ indicators, it is found that all indicators of the two universities have the same values except the first one (MEx) which value for UHIIC is less than this for UMV (Table 5.1). Yet, it's inconsistent with the fact that HRQ value for UHIIC is greater than this of UMV. So, the second step is to take row values that were computed from the aggregating formulas (Table IV) and one found that besides UID and RCU that have the same values for the two universities, MEx and StNI have greater values for UMV than for UHIIC (MEx (3; 2,5), StNI (8,5;8)) and a less one for StEx (4,75; 5,25). Once again, it can't be determined from this whose HRQ value is greater than the other because each indicator follows a different scale. Third step is about comparing standardized values that take into consideration the scale of each indicator and make the difference between the

values of the same indicators for the two universities comparable. In Table 6.1, it can be noticed that even if HRQ is better in UHIIC than in UMV, UMV is better in MEx and StNI with respective differences 0,62 and 0,35.

As a conclusion, one can say that HRQ in UHIIC is globally better than this of UMV. However, MEx and StNI in UHIIC have a low performance than those of UMV.

B. Comparing Universities According to the IS Component "Data"

USMBA is the best university in term of DQ among the twelve Moroccan public universities according to IS managers, with a score of 0,95, while USMS has the lowest DQ score (-1,48). As it's already done for HRQ, one take the case of two universities that have scores close to the mean: UIZ (0,11) and UAE (-0,10), and try to pursue steps to an inclusive comparison. First of all, taking the qualitative attributes of DQ indicators (Table 5.2.), it's found that two out of four indicators have the same values for the two universities. For the other ones, Updating and Back up (UpBp) value is greater in UIZ than in UAE (Short, Depending on data sensitivity and application type) contrary to Relevance (RI) value which is less in UIZ than in UAE. The contrast between the two universities doesn't allow a global comparison of DQ. So, the second step is about taking row values that were computed

from the aggregating formulas (Table IV) and it's found that besides Str and LR that have the same values for the two universities, UpBp has a greater value for UIZ than for UAE UpBp (3; 1) and a less one for RI (1; 2). Once again, it can't be determined from the former whose DQ value is greater than the other because each indicator has a different range. Third step is about comparing standardized values that take into consideration the range of each indicator and make the difference between the values of the same indicators comparable. In Table 6.2, it is noticed that even if DQ is better in UIZ than in UAE, UAE is better in UpBp with a difference of 0,26.

As a conclusion, one can say that DQ in UIZ is globally better than this of UAE. Nevertheless, RI in UIZ has a low performance than this of UAE.

VI. CONCLUSION

This study makes two important contributions to research on Information System Quality. The first novel aspect of this model is that it allows having numerical values for all indicators instead of qualitative description. These numerical values contribute to give each IS component standardized values able to provide an objective measure and an unbiased comparison.

Second, the findings provide scaled values of all the model indicators and components, thus enabling to arrange IS component of an organization from the lowest to the highest component performance. Thereafter a particular attention is given to the components with less performance level and go down to the indicators that compose them in order to highlight those with low values. Here one can point out precisely the weaknesses of ISQ in the organization, and can therefore propose corrective measures.

The data used in our research are collected from IS managers while data required to complete the whole picture of ISQ on an organization is from all IS intervening who are in addition to IS managers, technical staff, functional staff and users. As future research that are partly underway, once data from all IS intervening are collected, the same steps of the present study will be followed, leading to numerical values of all ISQ indicators and components. Thereafter, an aggregation of all IS intervening perspectives must be performed by organization entailing this way, one summarized value of ISQ for a specific organization that permits objective comparison.

REFERENCES

- [1] Aouhassi, S., & Hanoun, M. (2015). Information System Quality: State of the Art and New Model. *International Journal of Engineering Research and Technology*, 4(03), 589–594.
- [2] Aouhassi, S., & Hanoune, M. (2018). Information system qualification by component. In *ACM International Conference Proceeding Series (Vol. Part F1353)*. <https://doi.org/10.1145/3178461.3178478>
- [3] Arthur, J. D. (n.d.). Assessing the Adequacy of Documentation Through Document Quality Indicators*.
- [4] Azeroual, O., Gunter Saake, B., Jürgen Wastl, B., Azeroual Azeroual, O., Gunter Saake, dzhweu, & Wastl JuergenWastl, J. (n.d.). Data measurement in research information systems: metrics for the evaluation of data quality. *Scientometrics*. <https://doi.org/10.1007/s11192-018-2735-5>
- [5] Bakota, T., Beszédes, Á., Ferenc, R., & Gyimóthy, T. (2008). Continuous Software Quality Supervision Using SourceInventory and Columbus, 1–2.
- [6] Chawla, S. (2013). Review of MOOD and QMOOD metric sets, 3(3), 448–451.
- [7] Cyra, L., & Gorski, J. (2011). {SCF} — A framework supporting achieving and assessing conformity with standards. *Computer Standards & Interfaces*. <https://doi.org/http://dx.doi.org/10.1016/j.csi.2010.03.007>
- [8] Dimitrios, N. K., Sakas, D. P., & Vlachos, D. S. (2013). The Role of Information Systems in Creating Strategic Leadership Model. *Procedia - Social and Behavioral Sciences*, 73, 285–293. <https://doi.org/10.1016/j.sbspro.2013.02.054>
- [9] Gencel, C., Petersen, K., Ahmad, A., & Imran, M. (2013). The Journal of Systems and Software A decision support framework for metrics selection in goal-based measurement programs : GQM-DSFMS. *The Journal of Systems & Software*, 86(12), 3091–3108. <https://doi.org/10.1016/j.jss.2013.07.022>
- [10] Gorla, N., & Lin, S.-C. (2010). Determinants of software quality: A survey of information systems project managers. *Information and Software Technology*, 52(6), 602–610. <https://doi.org/10.1016/j.infsof.2009.11.012>
- [11] Heidari, F., & Loucopoulos, P. (2014). Quality evaluation framework (QEF): Modeling and evaluating quality of business processes. *International Journal of Accounting Information Systems*, 15(3), 193–223. <https://doi.org/10.1016/j.accinf.2013.09.002>
- [12] Letouzey, J.-L., & Coq, T. (n.d.). The «SQALE» Analysis Model An analysis model compliant with the representation condition for assessing the Quality of Software Source Code.
- [13] Letouzey, J. L., & Ilkiewicz, M. (2012). Managing technical debt with the SQALE method. *IEEE Software*, 29(6), 44–51. <https://doi.org/10.1109/MS.2012.129>
- [14] Marinescu, R. (n.d.). Detection Strategies: Metrics-Based Rules for Detecting Design Flaws.
- [15] Mordal-Manet, K., Laval, J., Ducasse, S., Anquetil, N., Balmes, F., Bellingard Laurent Bouhier, F., ... McCabe, T. J. (n.d.). An empirical model for continuous and weighted metric aggregation.
- [16] Nagendra, A., & Deshpande, M. (2014). Human Resource Information Systems (HRIS) in HR Planning and Development in Mid to Large Sized Organizations. *Procedia - Social and Behavioral Sciences*, 133, 61–67. <https://doi.org/10.1016/j.sbspro.2014.04.169>
- [17] Nieves Pérez-Aróstegui, M., Bustinza-Sánchez, F., & Barrales-Molina, V. (2015). Exploring the relationship between information technology competence and quality management. *Cuadernos de Economía y Dirección de La Empresa*, 18, 4–17. <https://doi.org/10.1016/j.brq.2013.11.003>
- [18] Rothenberger, M. A., Kao, Y.-C., & Wassenhove, L. N. Van. (2010). Total quality in software development: An empirical study of quality drivers and benefits in Indian software projects. *Information & Management*, 47, 372–379. <https://doi.org/10.1016/j.im.2010.10.001>
- [19] Sarrab, M., & Hussain Rehman, O. M. (2014). Empirical study of open source software selection for adoption, based on software quality characteristics. *ADVANCES IN ENGINEERING SOFTWARE*, 69, 1–11. <https://doi.org/10.1016/j.advengsoft.2013.12.001>
- [20] Shah, M. (2014). Impact of management information systems (MIS) on school administration: What the literature says. *Procedia - Social and Behavioral Sciences*, 116, 2799–2804. <https://doi.org/10.1016/j.sbspro.2014.01.659>
- [21] Suzila Kassim, E., Fatiany Abdul Kader Jailani, S., Hairuddin, H., & Hamiza Zamzuri, N. (2012). Information system acceptance and user satisfaction: The mediating role of trust. *Procedia -Social and Behavioral Sciences*, 57, 412–418. <https://doi.org/10.1016/j.sbspro.2012.09.1205>
- [22] Zahedi, F. (2003). Quality Information Systems. *Encyclopedia of Information Systems*, 3, 631–646. <https://doi.org/http://dx.doi.org/10.1016/B0-12-227240-4/00145-3>
- [23] Zheng, Y., Zhao, K., & Stylianou, A. (n.d.). The impacts of information quality and system quality on users' continuance intention in information-exchange virtual communities: An empirical investigation. <https://doi.org/10.1016/j.dss.2012.11.008>

Using Fuzzy Clustering Powered by Weighted Feature Matrix to Establish Hidden Semantics in Web Documents

Dr.Pramod D Patil, Dr.Parag Kulkarni
Department of Computer Engineering
Dr. D.Y. Patil Institute of Technology
Pune, INDIA

Abstract—Digital Data is growing exponentially exploding on the 'World Wide Web'. The orthodox clustering algorithms obligate various challenges to tackle, of which the most often faced challenge is the uncertainty. Web documents have become heterogeneous and very complex. There exist multiple relations between one web document and others in the form of entrenched links. This can be imagined as a one to many (1-M) relationships, for example, a particular web document may fit in many cross domains viz. politics, sports, utilities, technology, music, weather forecasting, linked to ecommerce products, etc. Therefore, there is a necessity for efficient, effective and constructive context driven clustering methods. Orthodox or the already well-established clustering algorithms adhere to classify the given data sets as exclusive clusters. Signifies that we can clearly state whether to which cluster an object belongs to. But such a partition is not sufficient for representing in the real time. So, a fuzzy clustering method is presented to build clusters with indeterminate limits and allows that one object belongs to overlying clusters with some membership degree. In supplementary words, the crux of fuzzy clustering is to contemplate the fitting status to the clusters, as well as to cogitate to what degree the object belongs to the cluster. The aim of this study is to device a fuzzy clustering algorithm which along with the help of feature weighted matrix, increases the probability of multi-domain overlapping of web documents. Over-lapping in the sense that one document may fall into multiple domains. The use of features gives an option or a filter on the basis of which the data would be extracted through the document. Matrix allows us to compute a threshold value which in turn helps to calculate the clustering result.

Keywords—Fuzzy; clustering; web document; feature matrix

I. INTRODUCTION

Let us now try to understand the need or motivation of the system. With an incredible circulation of several hundred million sites worldwide, the ever changing cluster of documents over the internet is getting bigger and bigger every day. This incorporates some very important and as well very difficult challenges. Over the preceding duration of ten years there has been incredible growth of data on World Wide Web. It has become a major source of information. Internet web generates the new defies of information retrieval [10] as the amount of data on web as well as the number of users using web growing rapidly. It is challenging to quest through this tremendously large catalogue for the information desired by

user. Also the traditional clustering algorithms like the k-means, probabilistic algorithms, k-medoid, and density based algorithms, constraint based algorithms and hierarchical algorithms fail to generate a result which render or convey the cross linked relations between the web documents. The other most important aspect was the traditional clustering algorithms use the standard numpy arrays which are very slow and not so effective in time complexity wise processing. Also, these traditional clustering algorithms face the issue of 'Concentration Measure' or 'Curse Dimensionality'. This was the motivation to propose a new algorithm using Weighted Matrix applying the Fuzzy Logic method. This would suffice the end user queries correctly. As explained earlier the amount of information on web is exponential and be termed as information burst, there is critical need to device the system that renders correct classification of data and should fetch correct result to the end user. Let us have a detail overview of components of our system and let us understand what operations it is designated to do.

II. RELEVANT TERMS AND DEFINITIONS

In this section, the relevant terms, tools, data mining process and techniques which are required for successful implementation of the experimental setup. Let us now start with the Web Crawlers. Search Engines use crawlers to collect data and then store it in database maintained at search engine side. For a given user's query the search engines searches in the local database and very quickly displays the results. The entire Knowledge Discovery System is shown in Fig. 1.

A. Web Crawlers

Web crawling is an imperative method for amassing data on, and custody up with, the speedily intensifying Internet. Web crawling can likewise be baptized as a graph search problem as web is considered to be a large graph where nodes are the pages and edges are the hyperlinks. Web crawlers can be used in various areas, the most prominent one is to index a large set of pages and permit other people to search this index. A Web crawler does not really move all over the place on the computers linked to the Internet, as viruses or bot agents do, as a substitute it only directs entreaties for documents on web servers from a set of already sites. However the web crawlers have progressed, there has remained significant weakness in search engines outstanding to the complex, inter related (linked or cross domain documents) in the document

assembly. Polysemies, synonyms, homonyms, phrases, dependencies and spam's act as hindrance to the search engines and therefore hampering the results returned. Also the vagueness or irrelevance of the user probes increases the ambiguous results fetched [2].

B. Pre-processing

Data pre-processing as shown in Fig. 2 exists an often neglected step but very important and is of prime importance since data pre-processing forms the foundation step of additional analysis and dispensation of data. Data pre-processing involves following five steps:

1) Data Cleaning

This step has operations like to fill values which are missing, smoothen out the noisy data, detecting or eliminating outliers, and deciding discrepancies.

2) Data Integration

It involves integrating data using numerous databases, data cubes, or collections.

3) Data Transformation

In this step we perform normalization and aggregation operations on data which has been integrated from various data sources.

4) Data Reduction

In this step we condense the quantity of data and produce the similar investigative results.

5) Data Discretization

In this step of data preprocessing we perform discretization operations like replacing numerical attributes with nominal ones.

[1]. Data congregation approaches are often range values, irregular, missing values. Analyzing data that hasn't been properly processed, such data can produce misleading results. Thus, pre-processing is primarily important step formerly running an investigation. Data fetched using a web crawler needs significant amount of processing before it is fed to the 'Fuzzy Clustering Algorithm' (FCA). Data in actual world is unclean which means it is incomplete. Incomplete data means it lacks attribute or the data in which we are interested in. The second part of dirty data is that it contains noise. Noisy data means that there are inaccuracies or outliers in it. The third part of dirty data is that it is inconsistent. Inconsistent means that the facts are not in correct format or the data lacks proper coding and naming format. If there is no good quality of data available then the data that would be eventually loaded in the data warehouse would be of low standards. The mining algorithms would yield a junk result out of the data warehouse. For data to be in correct format for data mining it should possess some valuable qualities where the data mined would be of highest quality [7].

These desirable qualities are precision, reliability, comprehensiveness, attribute value and most importantly timeliness. The most vital part of Preprocessing is cleaning the data. If the right data is not fed in we cannot expect the right output. Therefore cleansing of data is most vital. Missing data means computing the missing values. Adding missing values means filling the missing values with the average value derived by mean method. It also has a step of removing the noisy data. As discussed above noisy data means the data which comprises errors or outliers. Data cleaning also involves removing of inconsistencies. Inconsistent data removal means removing the data which falls into outlier range. Data can be collected from multiple formats like different databases, different file formats. The utmost vital part is to collate this data and then cleansing this data. It involves converting the dates to one particular format, converting numeric data to proper decimal format. It similarly involves performing binning on the numeric data. Filling the missing data is done by usually adding a tuple or replacing the missing data by a global constant. Applying the data cleansing task in my work the primary step is to remove the stop words from the data which has been fetched by the web crawler [4, 8]. Elimination of stop words and stemming [11]: In this phase, data which has less semantic is removed. Meaning, the full stops, commas, conjunctions etc. are removed. The data of the web documents fetched by the web crawler is equated with a bag of words – Stop words. The matched records are eliminated from the data file. Stemming process is a pre-processing step making the data ready for the next step. It is very important in most of the Information Retrieval systems. The main perseverance of stemming is to decrease diverse grammar pertaining forms or the words like its noun forms, adjective forms, verb forms, adverb forms etc. to its root form. The goal of stemming is to diminish deviational forms and occasionally derived various formations of a word to a conjoint base form. The available data is now further processed [5, 6].

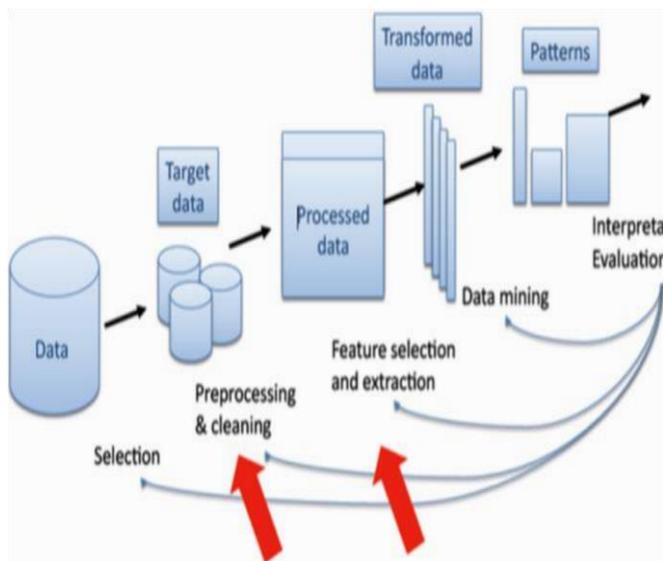


Fig. 1. Knowledge Discovery System [12]

The phrase – If you input the junk data that is ‘Garbage In’, then you be surely getting the junk output that refers to ‘Garbage Out’ is particular to the domain of machine learning

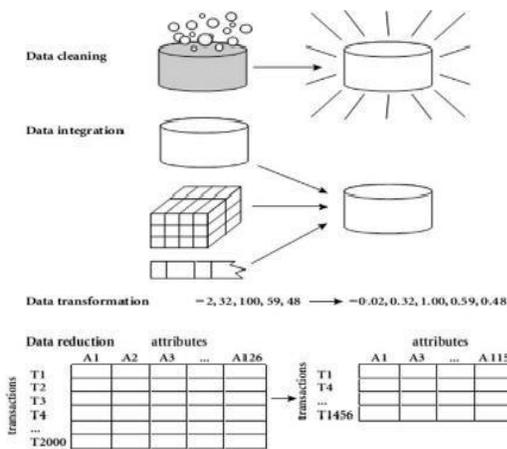


Fig. 2. Data Pre-Processing [12]

The next step of data pre-processing is integrating data from various databases, files, cubes together. Data integration means combining data into intelligible store. It also means schema integration. Schema integration means integrating metadata from different sources. It also includes identifying the attribute mismatch and performing a correcting action. This data is then fed to Data Reduction engine. In data reduction, the data which is redundant is removed. Data redundancy occurs because data is integrated from multiple databases, files etc. Meaning that same attribute might have been referred in a different way or the value of same attribute would have been derived or calculated. Redundant data is recognized by co relational analysis. Data reduction also involves performing numerosity reduction on data. It means to apply the linear regression model on data. This step is followed by Data Transformation. Data transformation means transforming the data in a format which is consistent throughout. It involves normalizing the data and aggregating it. The smoothing process of data transformation removes the noise from data. Aggregation step means aggregating the data into summarized cubes. Normalization activity means scaling the data such that it falls under particular range. It also includes construction of new attributes. It states that the data now has been fully transformed and ready to be loaded in the warehouse. We can conclude that data preparation is a critical issue for both data warehousing and data mining, as actual world data tends to be imperfect, noisy, and unpredictable. Data preparation involves data cleaning, data integration, data transformation, and data reduction. Data cleaning mechanism could be used to fill in missing values, lessen noisy data, detect outliers, and correct data inconsistency. Data integration loads data from multiples sources to form an intelligible data store. Metadata analysis, correlated data analysis, data skirmish detection, and the determination of semantic meanings add to smoothing the data. Data alteration techniques confirm the data into appropriate forms for mining. Data reduction methods such as dimension reduction data cube aggregation, numerosity reduction, data compression and discretization could be used to get a reduced depiction of the data, while minimizing the loss of information content. Concept hierarchies establish the attributes by the values or dimensions into measured levels of abstraction. They are methods of discretization that is predominantly

useful in multilevel mining. For numeric data, practices such as data segmentation by divider documentations, histogram analysis, and clustering analysis can be used [9].

C. Feature extraction

Data mining is the cumulative task of data analysis and detection algorithms to perform automatic extraction of information from vast amounts of data. This process bonds many practical areas, counting databases, human computer interaction, statistical analysis, and machine learning. A typical data-mining chore is to forecast an unidentified value of circa attribute of a new occurrence when the values of the supplementary qualities of the new occurrence are recognized and a collection of instances with known values of all the attributes is given. Most importantly in numerous applications, data is the subject of analysis and dispensation in data mining, is multidimensional, and presented by a number of topographies. There are moreover many dimensions of data that it is relevant to several machine learning algorithms and denote the extreme raise of computational complexity as well as classification error with data having high expanse of dimensions. Hence, the dimensionality of the feature space is habitually abridged afore cataloguing is commenced [3]. Feature extraction is one of the dimension measures for lessening techniques. Feature extracts a subset of novel features from the unique feature set by means of some functional mapping possessing as much information in the data as possible. Many of the definite world applications has numerous features those are used in an effort to safeguard accurate cataloguing. If all those features are used for buildup classifiers, then they function in high dimensions, and the learning process becomes complex, which leads to high cataloguing error. Therefore, there is a necessity to condense the dimensionality of the features of data before classification. The key objective of dimensionality reduction as shown in Fig. 3 is to convert the high dimensional data samples into the space of low dimensions such that the core information contained in the data is preserved. Once the dimensionality is reduced, it aids us to improve the heftiness of the classifier [11, 22].

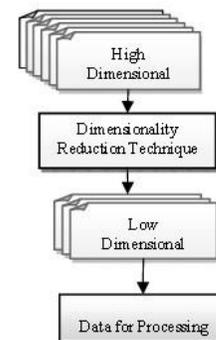


Fig. 3. Dimensionality Reduction Technique

Feature assortment is a technique to find good quality of germane features from the unique dataset using some data reduction and feature extraction measures. Feature extraction involves selection a feature, this is called as Feature Selection, Feature Selection step has turned out to be a thought provoking concern in the field of Pattern Recognition, Data

Mining, Machine Learning and Case Based Reasoning. Feature Selection is process of finding an ideal or suboptimal subset of 'n' features from the unique 'Features'. It requires a large search space to get the feature subset. The ideal feature subset is analyzed by evaluation criteria. The key objective of the feature selection is to decrease the amount of features and to remove the irrelevant, redundant and noisy data. Feature Selection includes various steps. These steps are portrayed in a diagrammatic state as below in Fig. 4.

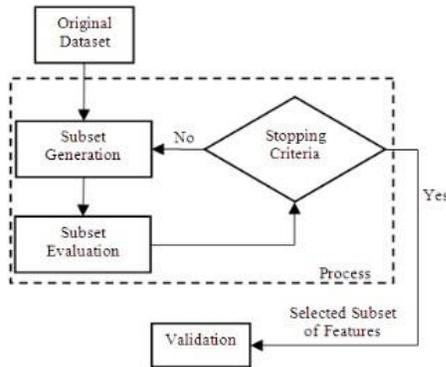


Fig. 4. Feature Extraction Engine

Feature selection mechanism is mostly classified into three types as shown in Fig. 5. They are, Filter Approach, Wrapper Approach and Hybrid Approach.

Feature selection method of 'Filtering an arithmetical measure used as a criterion for choosing the relevant features. This approach is calculated easily and very efficiently.

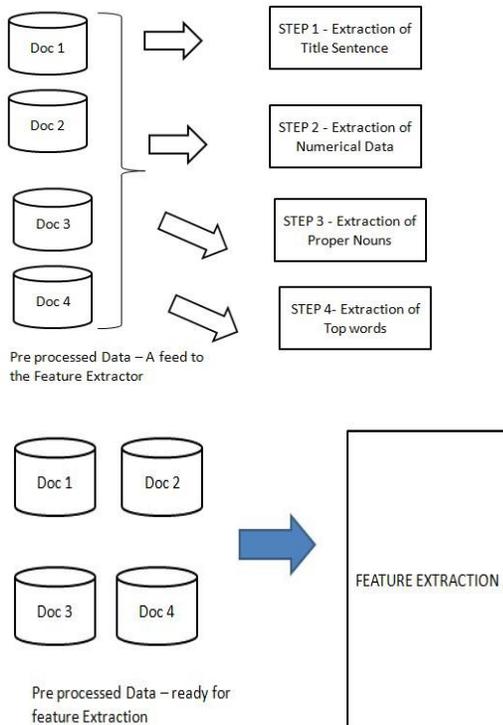


Fig. 5. Feature Extraction Process implemented in our work - A high level overview

With the processed data now available, we now extract the important 'features' available. The first step in feature extraction process is to fetch the 'Title Sentence'. The first line of the document is rendered as the 'Title Sentence'. The second step is extracting the numerical data in the data file. A scan is performed and all the numerical data present in the data files is counted. Next, all the nouns present in all the data files are listed. Only proper nouns are to be used. The last scan is done for the 'top words'. Each document now is scanned. The word whose count is highest is regarded as top word. A list of such words is made in descendant order rendering to the web documents.

D. Fuzzy Logic

Fuzzy Logic System are those which produce satisfactory but definite output in rejoinder to imperfect, vague, partial, or imprecise (fuzzy) input. Fuzzy Logic is a technique of perception that it is similar or resembles anthropological reasoning. The methodology of Fuzzy Logic tries to inherit the way of conclusion making in humans that encompasses all transitional possibilities between digital values Yes and No. The predictable logic that a system can comprehend takes exact input and gives a certain output as true or false, which is corresponding to human's YES or NO. The creator of fuzzy logic term, Lotfi Zadeh, detected that dissimilar computers, the human conclusion making embraces a range of likelihoods between YES and NO, such as: CERTAINLY YES, POSSIBLY YES, CANNOT SAY, POSSIBLY NO, CERTAINLY NO [13, 21].

Fuzzy logic contains of four vital phases: A Fuzzifier, Rule Base mapper, An Inference Engine and Defuzzifier.

Fuzzy Logic Systems Architecture is as follows:

1) Fuzzification Module

This unit alters the input to the systems, which are in the form of crisp numbers, into fuzzy sets. For example it transmutes the supplied crisp values to a linguistic variable by making use of the membership functions warehoused in the fuzzy knowledge base. Fuzzy linguistic variable is used to epitomize qualities straddling a particular spectrum or cross domain.

2) Fuzzy Knowledge Base Module

It stocks the conditions established on the If and then rules provided by experts. The fuzzy knowledge base is constructed on linguistic and membership functions.

a) Linguistic Variables

Linguistic variables act as input or output for the system. Their values are articulated in a natural language as an alternate to numerical values. A linguistic variable exists as a generally disintegrated into a group of linguistic terms.

b) Membership Functions

It is used for 'quantifying' the linguistic term. Membership functions are used in the fuzzification and defuzzification phase to plot the non-fuzzy variable as input to fuzzy linguistic terms as well as the reverse way round.

3) Inference Engine Module

It feigns the human cerebral method by creating fuzzy interpretation on the inputs and IF-THEN rules.

4) Defuzzification Module

It transmutes the fuzzy variable set gained by the corollary engine to a definite value [20].

The exponential growth of the Web has led to extensive expansion of web content. The enormous area of product data on the internet poses inordinate task to both users and online commerce. More users are turning towards online shopping because it is relatively convenient, reliable, and fast; yet such users usually experience difficulty in probing for merchandises on the internet due to information overload. Online selling has often been stunned by the rich data they have collected and find it challenging to endorse merchandises suitable to precise users. There is also the problem of futile consumption of the available huge amount of merchandise data from online transactions to support better decision making by both consumers and suppliers. To discourse these information overload problems, e-learning, e-commerce, e-newspapers data stores are now smearing mass customization ideologies not to the merchandises but to their staging in the online.

Fuzzy Logic System as shown in Fig. 6 could be understood as a system which maps nonlinear data as an input to a scalar output data set. Fuzzy sets obligate powerful decision making ability and hence attracted rising consideration and curiosity in recent IT, data generation method, decision building, pattern acknowledgement, and diagnostics and data analysis among others. When a problem has vibrant or evolving behavior, fuzzy logic is a suitable contrivance that deals with such problem. In short to say, fuzzy logic has m tier in providing precise solutions to problems

That encompasses the manipulation of numerous variables.

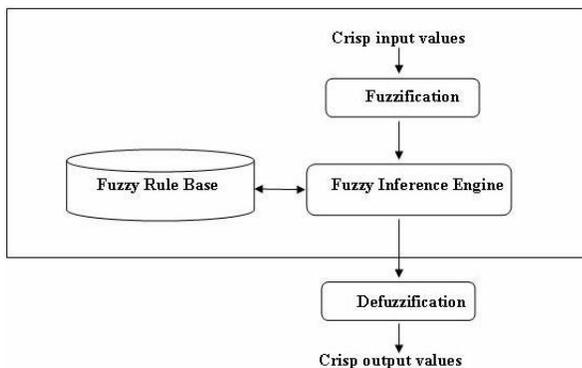


Fig. 6. The depiction of the Fuzzy Logic System [14]

The method of fuzzy logic systems are as follows:

- a) Define input and output crisp variables.
- b) Define the membership function.
- c) (iii) Convert crisp input data into linguistic fuzzy values, using membership function, called fuzzification.
- d) Evaluate the rules, using inference engine .

e) Construct the output crisp data, from fuzzy linguistic values, called defuzzification.

Fuzzy logic bids several unique features that make it a predominantly decent choice for many control problems.

a) It is inherently robust since it does not require precise, noise-free inputs and could be programmed to fail safely if a feedback sensor quits or is destroyed. The output regulator is a smooth control function notwithstanding an extensive assortment of input variations.

b) Since the Fuzzy logic checker processes the user-defined rules prevailing the target control system, it can be altered and tweaked easily to improve or radically alters system performance. New sensors can straightforwardly be fused into the system merely by engendering apt governing rules.

c) Fuzzy logic is not restricted to a few feedback inputs and control outputs, nor is it essential to measure or calculate rate-of-change restrictions in order to implement. This permits the sensors to be economical and imprecise thus keeping the inclusive system cost and intricacy low.

d) Due to the rule-based process, any equitable number of inputs can be administered and numerous outputs engendered.

e) Create Fuzzy logic membership functions that express the implication (values) of Input / Output relationships used in the rules.

Fuzzification is conversion of crisp variables into linguistic variables, and it is the central unit for fuzzy logic system. Variable pertaining to linguistic sense such as age might obligate a value such as 'young' or 'old'. However, the noteworthy efficacy of linguistic variables is that they can be amended via linguistic verges pragmatic to primary terms. Prof. Zadeh has recommended the notion of fuzzy variables. Although variables in arithmetic typically gross data which in the form of numbers, if the data which is not numeric then linguistic variables are often used to simplify the countenance of rules and facts. The usage of linguistic variables in numerous applications cuts the overall computation complexity of the application. Linguistic variables obligate to be predominantly useful in complex non-linear applications.

III. LITERATURE SURVEY

Ojokoh et al. in their work in the paper titled "A Fuzzy Logic Based Personalized Recommender System" [15] communicates to apply Fuzzy logic algorithms to the e-commerce space to drill to exact customer requirement. They carried out the experiments using laptops of various brands and configurations that customers usually search on various e-commerce sites. It defines the Fuzzy near compactness concept is engaged to measure the resemblance between customer needs and merchandise features. The ever-increasing figure of E-retail, e-commerce websites on the internet has led to data overload with over hundreds and thousands of customers. So it is challenging for customers of certain merchandises to discover information regarding merchandises in an attempt to purchase products that best satisfies them. This has led in reduction of the amount of product sales in the

e-commerce domain. The work in this paper highlights a personalized recommender system motivated by fuzzy logic method. The offered system intelligently mines data about the features of laptop computers and offers professional services to potential consumers by endorsing ideal merchandises grounded on their distinct requirements. They measured the result of the offered system by means of fifty laptop computers brands and configurations from Acer, HP, Sony, Dell and Toshiba. We studied the Fuzzy Logic implementation done in this paper. We also got to know how large data sets can help in an efficient fuzzy classification.

ChrisCornelius, Jie Lu et al. in their paper titled “*One and Only Item recommendation with Fuzzy Logic Techniques*” [16] implement a Collaborative Filtering method which is the abstract framework for endorsing one and-only items. It practices fuzzy logic, which permits to reflect the graded/uncertain data in the domain, and to range the CF paradigm, overcoming limitations of existing practices. The conceivable use of this Collaborative Filtering is in the e-government application. There is a personalization of e-government facilities intended at custom tailoring the content government made available to the end user. In several countries, e-government applications are increasing speedily and the quantity of e-government websites, as well as the assets and services provided, are dynamically increasing. This has caused a delinquent wherein citizens may find it more and tougher to locate relevant data from these websites. Matching specific citizens and businesses interests and needs is therefore one of the main trials for e-government services, and intelligent decision support. This paper gave us the idea of a fuzzy framework where a recommender system was constructed. It also gave an idea to construct a fuzzy algorithm which can be generically applied to all the cross domains.

Andreas Meier, KuisTeran in their work titled “*A Fuzzy Recommender System for eElections*” [17] describe the recommender system which is grounded on fuzzy logic and fuzzy clustering mechanism. It related to construction of an architecture for recommender system which can be used in e-Democracy and e-Elections applications. The use of this system enhances and succor voters in making verdicts by providing data about contenders close to the voter’s preferences and tendencies. The usage of recommender systems for e-Government is used to decrease data overload, which might help to advance self-governing processes. Fuzzy clustering investigation differs from classic clustering where the interpretations belong to only one cluster. Moreover, classic clustering makes no use of plodding membership. The recommender system approach fluctuates from collaborative filtering. The later one is built on historical experiences. It is suitable in the one and only scenario where events such as voting and election processes occur only once. This paper was a crucial reference as it implemented a filtering based fuzzy clustering technique.

Tung-Cheng, T-zone-I Wang et al. in their work titled “*A Fuzzy Logic based Personalized Learning System*” [18] shed light on use of fuzzy logic clustering the e-Learning domain. It employs fuzzy insinuation mechanisms, reminiscence cycle updates, apprentice preferences and systematic hierarchy process. The system has been used to cram any language. By

using fuzzy corollaries and personal reminiscence cycle updates, it is possible to find an editorial best suited for both a learner’s ability and their need to review vocabulary. After reading an article, a test is instantaneously provided to enhance a learner’s reminiscence for the words newly learned in the editorial. The methodology uses a questionnaire to realize a learner’s predilections and then uses fuzzy inference to find editorial of suitable exertion levels for the learner. It then employs review values to compute the fraction of editorial vocabulary that the learner must evaluate. It has cartels these three parameters to establish the article’s suitability formulae for computing the suitable level of articles for the learner. It uses memory to update the words so that the person seeking learning learns for the first time and also the words that appear that need to be reviewed based on the learner’s learning feedback. The consequences of these experiments vitrine that with intensive reading of pupilages as recommended by the approach, student can reminisce together new words and the words learnt in past easily and for longer time, thus competently enlightening the vocabulary ability of the learner.

A research conducted by JieZang in the field of a *Social Media based Personalized Recommender system* [19] based on Fuzzy logic describes a recommender systems which are built on intelligent computational abilities. From the topical past with the rise of data balloon on the internet, there is a consistent demand for the data processing engine for solving the problem of information overloading and information filtering. Present-day recommender systems hitch context-awareness with the personalization to deal the most accurate endorsements about diverse merchandises, services, and possessions. However, such systems arise across the issues, such as cold start, sparsity, and scalability that lead to vague endorsements. Computational Intelligence means not only improve endorsement accuracy but also markedly mitigate the above-mentioned issues. Computational Intelligent system as based on practices, such as: (i) fuzzy sets (ii) Artificial Neural Networks (iii) Evolutionary Computing, (iv) Swarm Intelligence, (v) Artificial Immune Systems.

IV. PROPOSED AND IMPLEMENTED SYSTEM

The main motivation was to propose an algorithm which uses the efficiency of the Matrix and weighted features with an application of fuzzy logic. This is a first ever attempt to create a fuzzy weighted matrix to extract the features of the data and then to form the overlapping logical clusters. Here in this section we are giving comprehensive emphasis on the design of the system. Each and every stage of the offered system is well narrated here. Along with the elucidation the complete system is well presented using the architecture. We are proposing a new algorithm using Fuzzy matrix and by using the weighted methods. The complete system as shown in Fig. 7 is dissected in four steps as discussed below.

The process proposed is: A web crawler would fetch number of web documents and store them in a folder. A web crawler would mainly act as a Data Source or Data Collector in our project work. The data fetched from the web crawlers is fed as an input to the pre-processing engine. The pre-processing engine follows the pre-designed steps of data

cleansing and gives out a bag of words corresponding to the document as output. The pre-processing engine is devised with three algorithms. That means it is a three step process. The first step is to 'Remove the special symbols'. The second step is 'Removing the stop words' and the third step is 'Removing the Stemming and deriving the root form of the word'. The output of this step is then fed as input to the further module. This is then fed to the feature extractor. The Feature extractor extracts the features from the bag of words pertaining to web document. It extracts the semantic features such as Numerical Data, Nouns, Title Sentence and Highest occurring word. It is actually based on accepting the accuracy and number of top words constructs a matrix. A weighted feature matrix is build up and using fuzzy logic the overlapping structures of the web documents are revealed as a final output.

1) Web Crawler

A web crawler plays an important part in this project. It mainly acts as a Data Source / Data Collector. The web crawler would fetch in number of web documents and parse them using the open source Google parser and store them in a folder called as WEBPAGES_REPOSITORY. This data set of web documents is then used as an input to the Pre-Processing Engine.

2) Pre-Processing

Pre-processing is vital step in data mining systems as it condenses the scope of the data required for processing. This condensed size minimizes the cost and space complexity of the system as fewer quantities of data are needed to be processed.

We have devised three algorithms for Pro-processing module:

- Special Symbol Removal

A special symbol removal algorithm is devised. It scans the bag of words in the array list and removes the special symbols from it e.g.!,@,#,\$,% etc. These special symbols do not contribute in result generation; hence it is worth to remove all the special symbols.

Algorithm for special symbol removal

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for $i=0$ to N (Where N is length of V)
Step 5: for i^{th} word of N check for its occurrence in Special Symbols repository
Step 6: if present then remove the special symbols
Step 7: else return the remaining words
Step 8: stop

- Stop Words Removal

Stop words are the words used as a supporting word in content to bring the semantics in the sentence; however, after this discarding the meaning of the sentence is not changing too

much extent. Hence they are removed here by maintaining one repository for comparison. This repository contains the 500+ stop words.

Algorithm to find stop word

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for $i=0$ to N (Where N is length of V)
Step 5: for i^{th} word of N check for its occurrence in Stop words repository
Step 6: if present then remove the stop words
Step 7: else return remaining words
Step 8: stop

- Stemming

Stems are used to derive word. Generally the words are derived for making the correct use of tenses. Unnecessarily this stems increase the system costing hence they are removed over here. No stemming algorithm is there which gives 100% accuracy.

Algorithm for stemming

Step 0: Start
Step 1: Read string
Step 2: divide string into words on space and store in a vector V
Step 3: Identify the duplicate words in the vector and remove them
Step 4: for $i=0$ to N (Where N is length of V)
Step 5: for i^{th} word of N check for its occurrence in Stemming extensions repository
Step 6: if present then process the word back to its root form
Step 7: else do nothing
Step 8: stop

3) Feature Extraction

As data contains tons of features it's not worth to consider the complete content for the further operations. Feature extraction is essentials step in data mining. It is used for fetching the required data i.e. features from the huge set of data. In our proposed work four features are extracted.

- Title Sentence

Title sentences are the one which represents the first sentence of the file content. The reason behind this extraction is to give a proper name to the cluster because each cluster is named by the title sentences.

- Numeric Data

Numeric data plays vital role in file content as the most of the important data are represented using numerical values only. So by considering this thing we extracted numerical values from the file content.

- Proper nouns

Proper nouns are the words which represent the person or place. For extraction of this feature a dictionary is used. So to access this dictionary jxlapi offers all the necessary functionalities.

Algorithm to find noun

- Step 0: Start
- Step 1: Read string
- Step 2: divide string into words on space and store in a vector V
- Step 3: Identify the duplicate words in the vector and remove them
- Step 4: for i=0 to N (Where N is length of V)
- Step 5: for ith word of N check for its occurrence in Dictionary (Open source dictionary api used)
- Step 6: if present then return true
- Step 7: else return false
- Step 8: stop

- Top Words

Top words are the important words of the sentence. Here in this feature the frequency of the each word are found out. The word which repeat more time is needed to consider as it have the more weightage in the file content.

Algorithm to find Term weight words

- Step 0: Start
- Step 1: Read string
- Step 2: divide string into words on space and store in a vector V
- Step 3: Identify the duplicate words in the vector and remove them
- Step 4: for i=0 to N (Where N is length of V)
- Step 5: for ith word of N check for its frequency
- Step 6: Add frequency in List Called L
- Step 7: end of for
- Step 8: return L
- Step 9: stop

4) Master Matrix Creation

Here in this step all the extracted features are taken as an input. From these entire features a one matrix is created. This is inspired from 'Vector Space Model', VSM which is an algebraic model for representing documents. So, particular feature of each file is compared with the respective feature of the other file. In this way all the four features are compared with four features of other file. This comparison led to a score of each file with other file.

Matrix Creation Process:

A weighted matrix is built, feature values are calculated against the every document in following way as shown in Table I:

TABLE I. WEIGHTED FEATURE MATRIX CALCULATIONS (FEATURES EXTRACTED ARE: TOP WORDS, NUMBER DATA, PROPER NOUNS, TERM WEIGHT)

	Feature Extracted	D 1	D 2	D 3
		(T,N,P,Tw)	(T,N,P,Tw)	(T,N,P,Tw)
D 1	(T,N,P,Tw)	0		
D 2	(T,N,P,Tw)		0	
D 3	(T,N,P,Tw)			0
D n	(T,N,P,Tw)			

Fuzzy Logic

The generated score from matrix is taken as input. The smallest and biggest score is calculated. Exactly five ranges are calculated starting from smallest value and end to largest value. Now the score is assigned to each of the scores calculated in master matrix step by checking the occurrence of the score in these five ranges. Once score is calculate a threshold of 2 is set. The file having threshold more than 2 is added to cluster and discards the file which fails to satisfy the condition.

Algorithm for Document clustering using Fuzzy matrix Weighted method

- Input: Merged Feature vector Fv
- User Accuracy as Ua
- Output: Cluster Set C= {c1, c2, c3...cn}
- Step 0: start
- Step 1: create matrix M of length Fv
- Step 1: For i=0 to Fv length (for each row)
- Step 2: For j=0 to Fv length (for each column)
- Step 3: Fvr= element of one row
- Step 4: Fvc=element of one column
- Step 5: Compare features and get score as Sc
- Step 6: Average Score as Asc=Sc/4
- Step 7: add Average to matrix M
- Step 8: End Inner For
- Step 9: End Outer For
- Step 10: for every file in M'sRows if (Asc>=Ua) then add into cluster Ci
- Step 11: return cluster set C
- Step 12: Stop

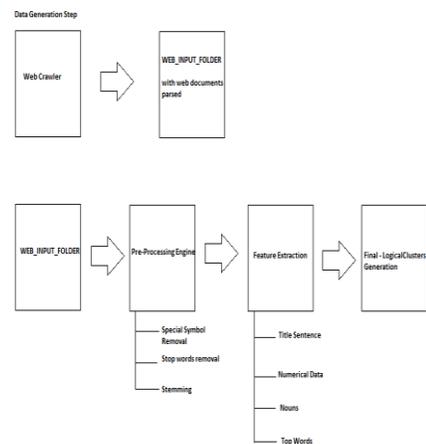


Fig. 7. Our Proposed System Architecture Overview

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

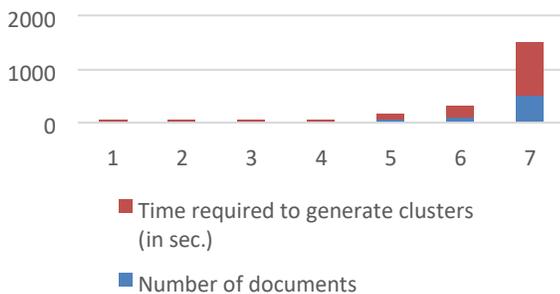
To show the efficiency of the system on experiment is conducted on java 1.6 based machine using Net beans as an IDE on windows machine having 2GB ROM and 500GB HDD. After doing the experiment by providing the files from different categories such as text, pdf, and doc the following observation is led.

We have considered the cross domain documents pertaining to various fields like Sports, medicine, finance, insurance, travel, music, etc. This data set was taken from the world renowned news channel which is an open source for data set and which is available to use for research work. The size of each document was about 2-3 MB text file. Furthermore we also provided the input as .mp3 files, .docx files, video files which were successfully handled and ignored by the system as currently we do not support these file types.

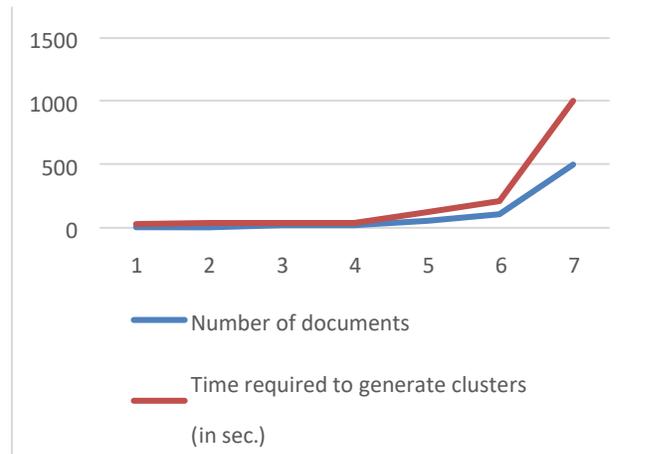
This system can be used as an overnight process, where you feed the system with huge volumes of documents and the system would successfully run and form the logical cluster groups of the documents. Logical clusters then would determine the relationships of various documents with each other.

TABLE II. TIME REQUIRED TO CLUSTER VS NUMBERS OF DOCUMENTS

Number documents of (Cross Domain)	Time (seconds) to generate overlapping clusters
5	30
10	34
15	41
20	47
50	117
100	209
500	1004



Graph 1: Performance measurement – Bar Graph representation of Number of Docs vs Time Required to Generate Clusters



Graph 2: Performance Measurement – Linear exponential depiction of Number of Documents to be clustered w.r.t Time Required

Graphs 1 and 2 signifies the clustering time. From the graph we can determine that as the numbers of documents increase exponentially the required time to generate the clusters marginally increases in folds as shown in Table II.

Application screen shots and Comparative study

Compared the Fuzzy clustering by Weighted Feature Matrix algorithm with the traditional clustering algorithms.

A comparative study was conducted using the Dataset obtained from the BBC news website. The data set comprised of documents pertaining to various domains like ‘Banks’, ‘Loans’, ‘Sports’, ‘Insurance’, ‘Weather’, ‘Politics’, ‘Music’, ‘Films’, ‘Geography’, ‘History’, ‘Literature’ etc. We found that K-Means, Hierarchical algorithms do not perform well due to the inability to recognize the semantic meaning of the document. Therefore there was a need to propose a new algorithm which you carry out the clustering task as per the hidden semantics meaning. Thus we have tried our best to propose a new algorithm which extracts the features from the web documents and then follows the weighted matrix for generating the logical clusters. The implemented results are depicted from Fig. 8 to Fig. 18. The login screen shown in Fig. 8.

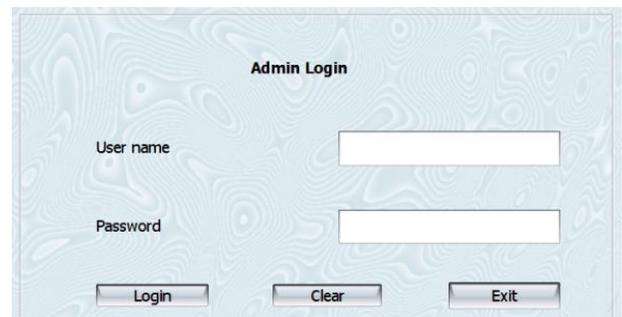


Fig. 8. Login Screen.

Now, Navigate to “System Settings”> Set “Accuracy”> Enter valid percentage from 0%-99% as shown in Fig. 9.

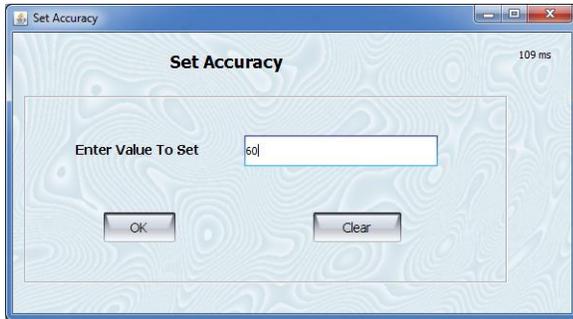


Fig. 9. Input Accuracy in %

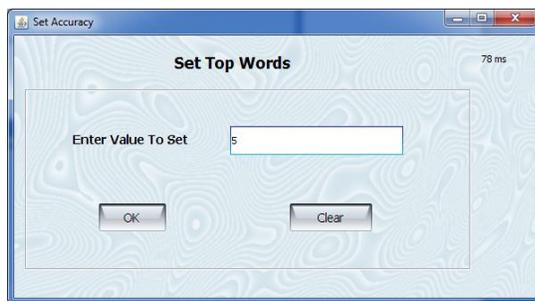


Fig. 10. Top Words

The process to set the value for Top words is depicted in Fig. 10. Now, navigate to “Folder Input”. Select the web pages repository which needs to be fed to our system as shown in Fig. 11.

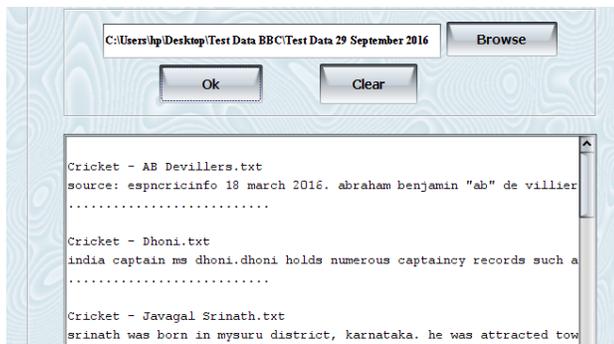


Fig. 11. Data Set Selection

Now, we will apply the “Pre-Processing” algorithms on the data imported from Web pages repository as shown in Fig. 12, 13 and 14.

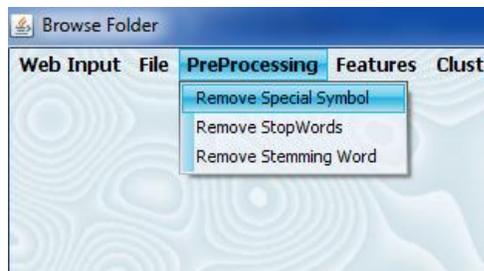


Fig. 12. Removal of special symbols

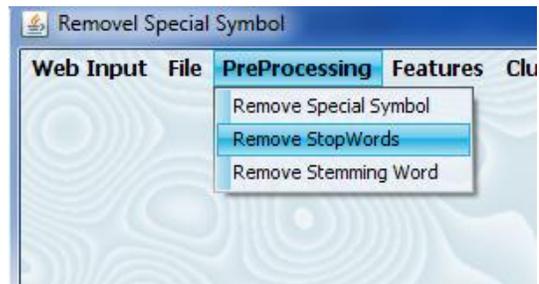


Fig. 13. Removal of stop words

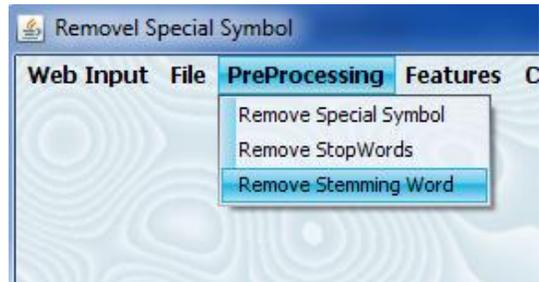


Fig. 14. Removal of stemming words

After the data has been pre-processed, we now feed this data to the “Feature Extractor engine”.

The following features are extracted in the given order as shown in Fig. 15, 16 and 17.

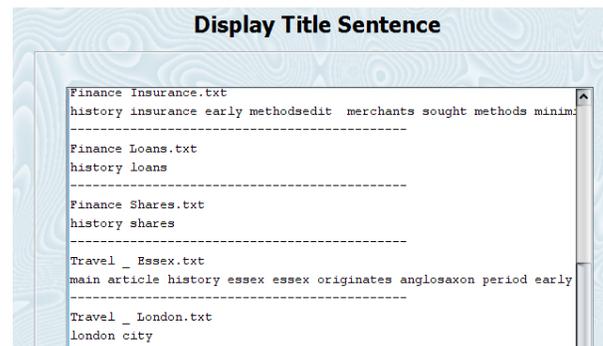


Fig. 15. Title Sentence

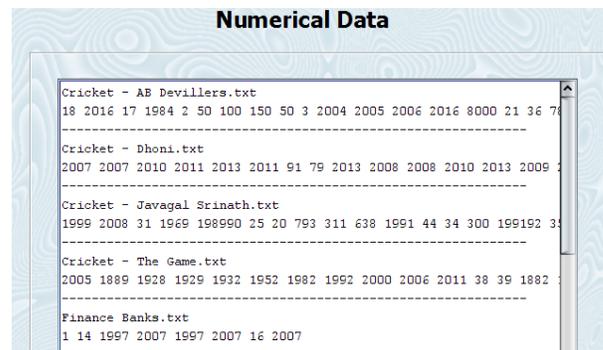


Fig. 16. Numerical Data

A dictionary scan is run, every word in the document is matched against the words in dictionary and only nouns are processed.

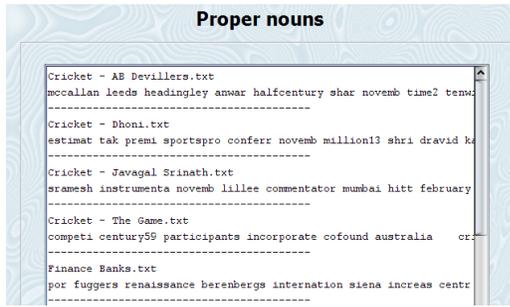


Fig. 17. Proper Noun

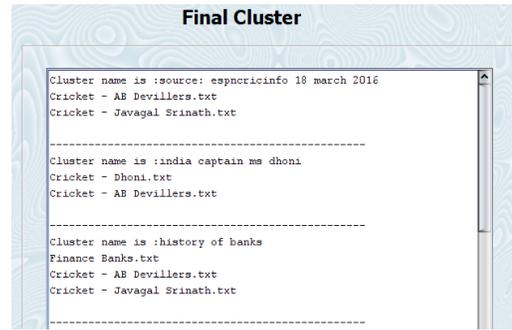


Fig. 18. Logical Clusters

Now, generate the Weighted Matrix

Cricket - AB Devillers.txt

[0, 0.14, 0.76, 0.24, 0.04, 0.0, 0.08, 0.0, 0.01, 0.19, 0.26, 0.02]

Cricket - Dhoni.txt

[0.64, 0, 0.42, 0.34, 0.06, 0.0, 0.02, 0.0, 0.02, 0.27, 0.16, 0.02]

Cricket - Javagal Srinath.txt

[0.47, 0.08, 0, 0.23, 0.03, 0.0, 0.1, 0.0, 0.03, 0.18, 0.25, 0.01]

Cricket - The Game.txt

[0.39, 0.12, 0.46, 0, 0.02, 0.0, 0.06, 0.0, 0.0, 0.18, 0.23, 0.02]

Finance Banks.txt

[0.91, 0.31, 1.44, 0.28, 0, 0.0, 0.22, 0.14, 0.14, 0.44, 0.06, 0.09]

Finance Derivatives.txt

[0.0, 0.0, 0.0, 0.0, 0.0, 0, 0.0, 0.0, 0.06, 0.0, 0.0, 0.0]

Finance Insurance.txt

[0.62, 0.06, 1.31, 0.06, 0.1, 0.0, 0, 0.02, 0.03, 0.29, 0.2, 0.15]

Finance Loans.txt

[0.0, 0.0, 0.0, 0.0, 0.15, 0.0, 0.12, 0, 0.12, 0.12, 0.0, 0.0]

Finance Shares.txt

[0.6, 0.07, 1.36, 0.05, 0.13, 0.05, 0.13, 0.12, 0, 0.18, 0.35, 0.0]

Travel _ Essex.txt

[0.43, 0.17, 0.62, 0.25, 0.05, 0.0, 0.13, 0.02, 0.03, 0, 0.34, 0.13]

Travel _ London.txt

[0.26, 0.04, 0.4, 0.15, 0.0, 0.0, 0.06, 0.0, 0.01, 0.2, 0, 0.06]

Travel_surrey.txt

[0.38, 0.09, 0.47, 0.1, 0.09, 0.0, 0.19, 0.0, 0.0, 0.38, 0.18, 0]

The final logical clusters formed are shown in Fig. 18.

TABLE III. EXPERIMENT CONDUCTED ON SYSTEM HAVING CONFIGURATION AS INTEL I7-970 PROCESSOR WITH 4 GB RAM HARD CLUSTER: DOCUMENT BELONGS STRICTLY TO ONE CLUSTER

	No. of documents as input	No. of Hard clusters	No. of overlapping clusters (Cross Domain)	Time required to form cluster (in seconds)
Proposed algorithm	120	0	32	157
K-Means algorithm	120	11	0	160

The number of clusters formed and time required in forming clusters using our proposed algorithm and K-means algorithm is shown in Table III. The proposed algorithm outperforms.

A set of documents used for evaluation has following features:

- 1) Number of documents per category
- 2) Evenness in number of documents in each category
- 3) Size of each document i.e. the number of words in each document
- 4) Similarity of documents of same category compared to similarity of documents of different categories.
- 5) Number of unique words in all the documents. The quality of the results of the clustering algorithms depends very much on the features of the set of documents on which it is applied. For example, some algorithms may give good results in case of large documents as compared to small documents.

The documents of BBC dataset are large news articles and thus the names of people, places, organizations, etc. play an important role in them and this gives the consideration of co-occurrence of words a huge importance. For example, 'Tendulkar' and 'cricket' are two different words which co-occur many numbers of times in news articles. Now, if an article contains only 'Tendulkar' then the feature based approach will still put it in the cluster of articles related to cricket or sports but this will not be the case with other algorithms.

VI. CONCLUSION AND FUTURE WORK

The experiment conducted shows that weighted feature matrix when combined with the application of Fuzzy Logic yields accurate results for the overlapping clusters of the web documents. Thus this gives us a deep insight of interlinked or

interconnected documents. It also gives us an option of clustering the web document by application of fuzzy logic by using the feature extraction method. Feature extractions enable the cardinality of the data that is to be extracted. We extracted proper nouns, numerical data, top words, term weight for this experiment. By using weighted matrix it gave us the flexibility for the calculations and an ease of computing the results. Matrix formation process on the basis of the features extracted is a unique method and the threshold value would reap us the results of the overlapping clusters. Large sets of web documents which are inter-related could be classified into clusters by using this novel way by the application of the fuzzy logic. The algorithm devised in this work is at very rudimentary stage and there are many possibilities for improvements. Some of the work that can be done on it is elaborated in this section.

This work can be enhanced by application of the Natural language Processing. Documents which are in Marathi, Hindi, Chinese, and Japanese, etc. could be easily classified into the logical overlapping clusters with the application of the fuzzy logic algorithm. Also, another enhancement could be increasing the feature set as per the domain example a feature as a Date could be used in formation of forensic data or historical data. Similarly domain knowledge from various fields like Medicine, Sports, Financial Services, Space Research, Weather Reports, etc. could be applied in this experiment.

This work could be migrated to next level for deep learning and data analytics by migrating it to “R” programming for reaping the best results.

As of now, considering the limitations, we have used the open source APIS for the crawler, parser, dictionary, reading the data from excel files. These open source APIS have a limitation on the volume of data that could be processed, the open source dictionary also does not cater all the noun forms of the words. It may be enhanced by using self-developed APIS.

REFERENCES

- [1] Kotsiantis, S. B., D. Kanellopoulos, and P. E.Pintelas. "Data pre-processing for supervised learning." *International Journal of Computer Science* 1.2 (2006): 111-117.
- [2] The wall, Mike. "A web crawler design for data mining." *Journal of Information Science* 27.5 (2001): 319-325.
- [3] M. Ram swami and R. Bhaskaran, "A Study on Feature Selection
- [4] Liu, Jin-Hong, and Yu-Liang Lu. "Survey on topic-focused Web crawler." *Appl. Res. Computer* 24.2629 (2007)
- [5] Frakes, William B. "Stemming algorithms." (1992): 131-160.
- [6] Govind MurariUpadhyay, KanikaDhingra, "Web Content Mining: Its Techniques and Uses", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 11, November 2013, pp.610-613
- [7] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996)
- [8] Munk, Michal, JozefKapusta, and Peter Švec. "Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor." *Procedia Computer Science* 1.1 (2010): 2273-2280.
- [9] Khasawneh, Natheer, and Chien-Chung Chan. "Active user-based and ontology-based web log data pre-processing for web usage mining." *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*.IEEE Computer Society, 2006.
- [10] JH Paik, MandarMitra, Swapan K. Parui, KalervoJarvelin, "GRAS: An effective and efficient stemming algorithm for information retrieval", published in *ACM Transaction on Information System (TOIS)*, Volume 29 Issue 4, December 2011, Chapter 19, page 20-24
- [11] M. Bacchin, N. Ferro, and M. Melucci 2005."A probabilistic model for stemmer generation". *Inf. Process. Manage.* 41, 1, 121–137.
- [12] *The Text Book of Data Mining*, Kimball
- [13] Subramanian Appavu Alias Balamurugan, Ramasamy Rajaram "Effective and Efficient Feature Selection for Large-scale Data Using Bayes' Theorem", *International Journal of Automation and Computing*, Volume6, Issue 1, Feb 2009, pp. 62-71
- [14] H. Ying, "A Fuzzy Systems Technology: A Brief Overview" IEE Press, 2000
- [15] Ojokoh, B. A., Omisore, M. O, Samuel, O. W, and Ogunniyi, T. O. Department of Computer Science Federal University of Technology, A Fuzzy Logic Based Personalized Recommender System, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.5, October 2012 1008
- [16] Chris Cornelis a, Jie Lu b , XuetaoGuo b , Guanquang Zhang,One-and-only item recommendation with fuzzy logic techniques, *Information Sciences*, Volume 177, Issue 22, 15 November 2007, Pages 4906-4921
- [17] Luis Ter´an and Andreas Meier Information Systems Research Group, University of Fribourg, A Fuzzy Recommender System for eElections, K.N. Andersen et al. (Eds.): EGOVIS 2010, LNCS 6267, pp. 62–76, 2010. c Springer-Verlag Berlin Heidelberg 2010
- [18] Tung-Cheng Hsieh, Tzone-I Wang*, Chien-Yuan Su and Ming-Che Lee , A Fuzzy Logic-based Personalized Learning System for Supporting Adaptive English Learning, January 2012, *Educational Technology & Society*
- [19] Aaditeshwar Seth and Jie Zhang School of Computer Science University of Waterloo, ON, Canada, A Social Network Based Approach to Personalized Recommendation of Participatory Media Content, Copyright c 2008, Association for the Advancement of Artificial Intelligence
- [20] I-Jen Chiang, Member, IEEE, Charles Chih-Ho Liu, Yi-Hsin Tsai, and Ajit Kumar, Discovering Latent Semantics in Web Documents using Fuzzy Clustering, *IEEE Transactions on Fuzzy Systems* (Volume: 23, Issue: 6, Dec. 2015), Page(s): 2122 - 2134
- [21] A. Ali, and N. Mehli, "A Fuzzy Expert System for Heart Disease Diagnosis", *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1, Pp. 134-139, 2010.
- [22] Khalid, Samina, Khalil Tehmina,NasreenShamila, A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning, *IEEE Science and Information Conference*, 2014, pp. 372-378.

Declaration: "The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper."

A New E-Health Tool for Early Identification of Voice and Neurological Pathologies by Speech Processing

Bouafif Lamia

ISTMT, LITI laboratory, Tunis Manar University,
2092 Tunis, Tunisia

Ellouze Noureddine

LITI laboratory, Tunis Manar University,
ENIT, BP 37, Tunis belvedere 1002, Tunisia

Abstract—The objective of this study is to develop a non-invasive method of early identification and classification of voice pathologies and neurological diseases by speech processing. We will present a new automatic medical diagnosis tool which can assist specialists in their medical diagnosis. The developed strategy is based on speech acquisition of the patient followed by audio features extraction, training and recognition by using the HTK toolkit. The computed parameters are compared to standard values from a codebook database. The experiments and tests are conducted by using the MEEI pathological database of KEY Pentax. The obtained results give good discrimination with a mean pathology recognition ratio about 95%. Finally, this E-Health application is helpful for the prevention of specific diseases and improving the quality of patient care as well as reducing the costs of healthcare.

Keywords—E-Health; voice disorder; HMM classification; feature extraction; MFCC; pathology recognition rate

I. INTRODUCTION

Developments of non-invasive methods for voice pathology diagnosis were developed in order to assist medical staff and otolaryngologists to conduct objective and efficient diagnosis. At present a number of classic diagnostic tools are available on the market which were based on speech measurements and imaging analysis.

Many studies which used the speech features extraction succeeded to obtain acceptable discrimination ratio between normal and pathologic speakers. Some of them have achieved classification accuracies between 70% and 90% [1]. In fact, acoustic analysis allows estimating a large amount of long-term acoustic parameters such pitch, formants, jitter, shimmer, Amplitude Perturbation Quotient, Harmonics to Noise Ratio and Normalized Noise energy [2]. These features are very useful for characterizing speaker disorders especially if they are associated with MFCC or RASTA-PLP coefficients.

In other references, advances in speech processing have contributed to the identification of some neurological diseases (Parkinson, dyslexia, scleroses) by voice parameters analysis. The developed method is based on the determination of speech parameters of a speaker from a hardware interface and software for digital acquisition and processing of the speech signal [3]. In this research, we will develop a speech processing tool for clinical observation and detection of pathological. This interface is intended, not only for patients

but also for people currently using voice (singers, teachers). The methodology is very easy and is based on a voice recording. Then, the extracted speech parameters are applied as inputs to the famous HTK toolkit (HMM classifier). The results are compared with normal and pathological values for a detection and classification disease by using the famous MEII database [4].

II. RELATED WORKS

During the last decade, several digital methods of pathological identification from speech processing have been used for the classification and early identification of some diseases. These strategies can be classified into three categories:

- The first method is based on the extraction of acoustic parameters and the search of new descriptors and metrics of quality, distortion and voice irregularities such as MFCC, RASTA, LPC coefficients, Jitter, Shimmer and Harmonic ratio. The MFCC parameters were considered in other numerous studies, such as [5] and [6]. In [5] subjects with nodules, edema and unilateral vocal fold paralysis were analyzed with not encouraging accuracy results (78%), while in [6] patients suffering from spasmodic dysphonia were selected.
- The second one is based on machine learning techniques such as SVM and LDA. Among several machine learning techniques existing in literature, Support Vector Machine (SVM) has been widely used in voice signal processing such as the work of L.Godino [7] and S.N. [8] with accuracy ratio of 86%.
- The third is a statistical method which uses Hidden Markov Models (HMM) or Gaussian mixtures (GMM). It is based on learning and testing procedures for voice recognition and classification. The learning procedure constitutes the codebook (database of speech models and parameters), hence the testing procedure consists the audio real time acquisition and recognition step.

For example, Emary [9] uses GMM algorithm on a very small subset of the SVD database containing 38 pathological and 63 healthy voices in order to identify neurological disorders.

A. The Studied Voice and Neurology Pathologies

Vocal fold pathologies can be classified as physical, neuromuscular, traumatic and psychogenic diseases. They affect the voice quality. Several voices, neurological, organic or genetic diseases are associated with speech disorders and dysfunctioning [10]. In fact, a voice disorder can generate a language disorder causing degradation of the voice and its intelligibility. These disorders can be divided into next classes:

- *Dysphonia*: It can be considered as an abnormality of the speech production and quality or a paralysis or a kind of laryngitis.
- *Dysarthria*: Is a speech disorder related to paralysis or to poor coordination of the muscles involved in the articulation. This disease has a neurological origin and conducts to dyslexia disease for children.
- *Aphasia*: Is a language disorder due to a lesion of the cerebral cortex. The patient no longer includes the meaning of words or can no longer be expressed [10].
- Sclerosis is a neurological disease in which affects the brain of a part of the brain and spinal cord [11]. It causes muscle weakness, trouble with sensation, coordination and speaking [12].
- Parkinson and Alzheimer are neurological diseases which affects the brain controls then body mechanisms and articulations.

B. Speech Aspects

The speech signal is full of physiological and acoustic parameters. It can inform as about the identity of the speaker, its health and even its emotional state.

The speech is characterized by its variability in amplitude and phase and its non-stationary behaviour. It is the result of a convolution between a phonation (glottis source) and an articulation (vocal tract). The source is characterized by the pitch F_0 , yet the vocal tract is characterized by a formant structure which reflects the resonance of the vocal tract given by the formants (F_1, F_2, F_3, \dots) [12]. For example, Fig. 1 and 2 represent an illustration of the waveform, spectrogram, pitch and formants parameters of speech and music signals.

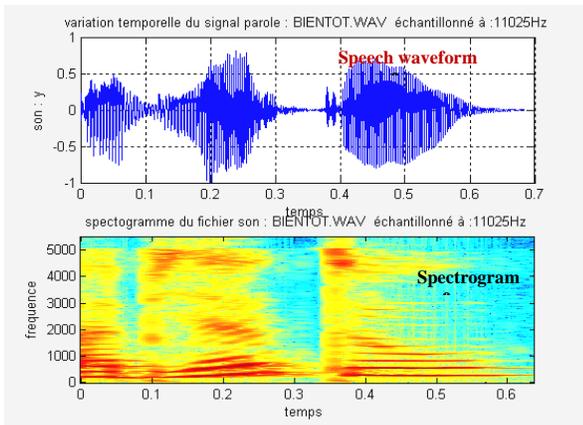


Fig. 1. Speech Waveform & Spectrogram.

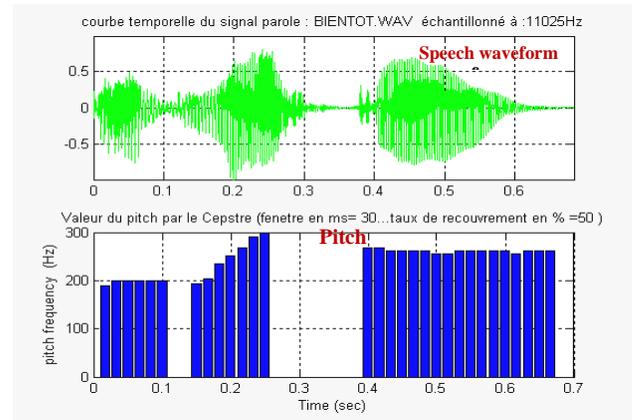


Fig. 2. Pitch Evolution of a Female Speech “Bientot.Wav” Sampled at $F_s=11025\text{Hz}$.

Fig. 1 and 2 represent the waveform, spectrogram and pitch profile of a female speech signal “bientot.wav” sampled at 11025Hz. According to Fig. 2, we can observe that mean pitch frequency is about 245Hz with a silent zone between 0.25 and 0.4 seconds.

The wide band spectrogram of Fig. 1 shows the formantic character of the speech illustrated by the red curves.

III. MATERIALS AND METHODS

In this work, we used the statistical HMM method, because it is very famous for speech recognition and synthesis and gives high accuracy for classification especially for high databases and noised environments. Other references used SVM, LDA and GMM classifiers [13, 14].

A. Speech Pathology Database

We have used the MEEI database of disordered voice (Kay Elemetrics Corporation) which was produced by the Kay Pentax [4]. The database is composed of many data dealing with the assessment of voice pathologies. It is considered as the most widely used dataset for research in pathological voice classification. The KAY database includes recordings of vowels pronounced by 53 normal subjects and 657 pathological voices coming from several diseases. The technical sheets are provided with the Recording and data files, containing information on the subjects (age, sex, language, smoking or not) and the results of the analysis calculated by the software MDVP.

TABLE I. CONTENT OF THE MEEI DATABASE

Pathology	Number of patients
Dysarthria	17
Gastric Refux	12
Ventricular compression	17
Nodule	6
Odeme	15
Disphonia	22
Normal	17

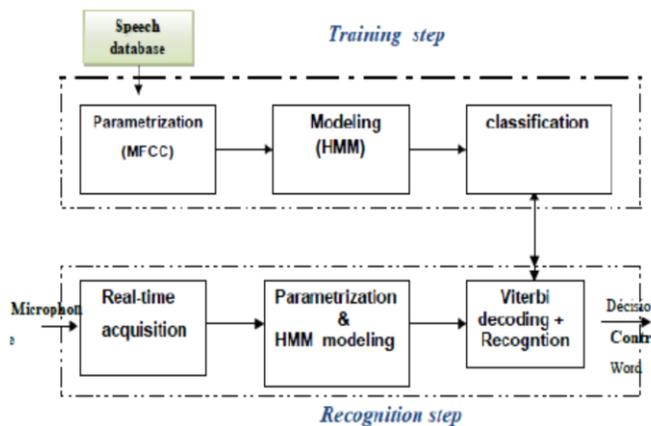


Fig. 3. The Training and Classification Method.

This software is also exclusively produced by Kay Pentax Corporation and is widely used in the clinical field as a tool for the recording and analysis of patient’s voice. The available pathologies are: the dysphonia, nodules, paralysis, polypoïde degeneration, and vocal cords disorder. These pathologies are recorded up to 10 seconds by men and women. Table I gives more details about the content of this database.

B. Pathology Identification with HMM

We have used the famous HTK platform based on the Hidden Markov Models (HMM) in order to recognize the pathological voices and a further disease classification. This

tool is a set of libraries and programs in C language developed at Cambridge University under the direction of Young in 1989 [15] in order to develop a performing technical Automatic Speech Recognition System. This toolkit is composed of:

- a speech database
- a training procedure by using the Baum-Walch and K-means algorithm for speech modeling. This step is applied on the speech database to constitute the reference codebook.
- a recognition procedure which is based on a real time acquisition and analysis , then a comparison with the training words by using Viterbi algorithm.

This procedure is illustrated by Fig. 3 where we can observe the different steps of parameterization (feature extraction), training, recognition and classification. In this step, the test audio model is compared with the codebook in order to find any similarity or coincidence with pathological models.

C. Speech Features Extraction

The first step of the speech analysis before modeling and coding is the parameterization of the speech frames into MFCC, LPC, PLP or RASTA coefficients. The Mel Frequency Cepstral Coefficients (MFCC) are the most famous method in speech processing, recognition and synthesis. Its principle is illustrated by Fig. 4.

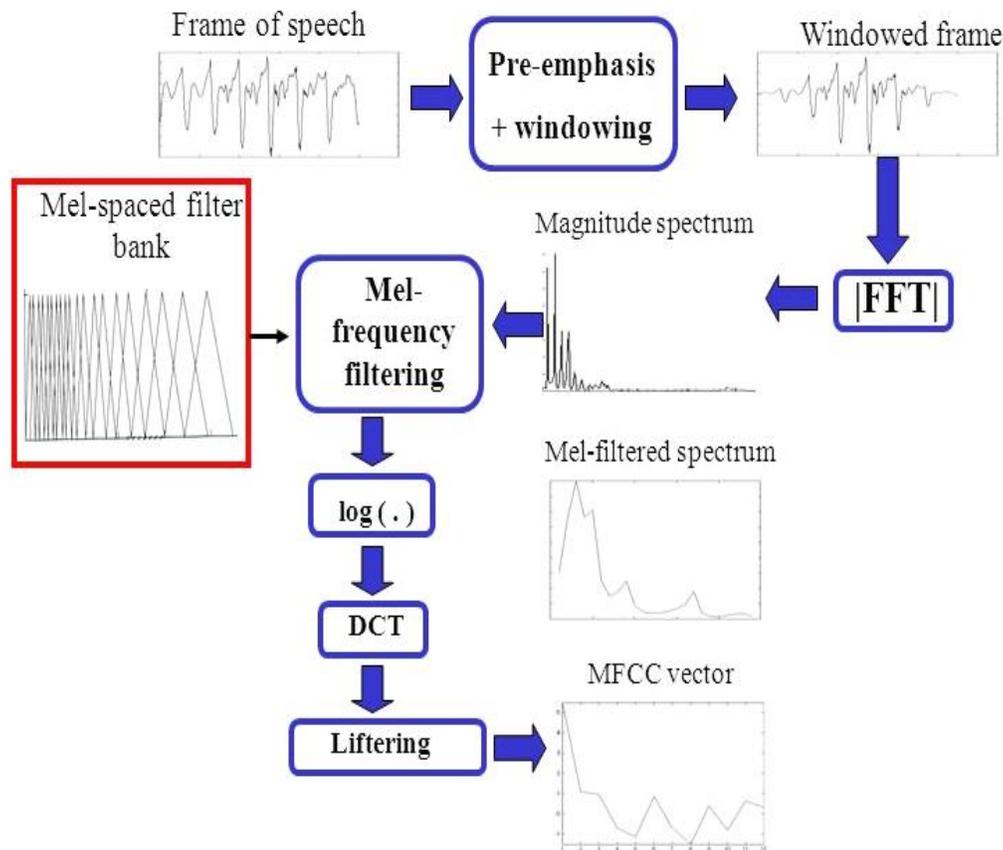


Fig. 4. MFCC Feature Algorithm.

In fact, MFCC is the most used for speech feature extraction and parameterization. MFCC algorithm which is represented by Fig. 2 can be expressed as [16]. The MFCC formula can be expressed as:

$$MFCC(i) = \sum_{k=1}^N \log(E_k) \cdot \cos\left[\frac{\pi \cdot i}{N} \left(k - \frac{1}{2}\right)\right] \quad (1)$$

Where: E_k : is the energy of the k^{th} filter

N : is the number of band-pass filters

Two others parameters are very useful in voice disorders analysis which is Jitter and shimmer. These indicators represent the irregularities and perturbations respectively in frequency and intensity. The expressions are given by next equations [16, 17]:

$$Jitter = \frac{\frac{1}{N-1} \sum_{k=1}^N |T_k - T_{k+1}|}{\frac{1}{N} \sum_{k=1}^N |T_k|} \quad (2)$$

$$Shimmer = \frac{\frac{1}{N-1} \sum_{k=1}^N |A_k - A_{k+1}|}{\frac{1}{N} \sum_{k=1}^N |A_k|} \quad (3)$$

With:

T : the pitch period

A : the amplitude of the pitch

N : the number of samples.

k : indices of the frame

D. HMM Training and Recognition

Fig. 5 represents the principle of the training-recognition-classification procedure. The training step uses a database or a codebook constituted of audio parameters and Baum-Walch and K-means algorithms. The recognition procedure uses HMM modeling and classification by using Viterbi decoder [18].

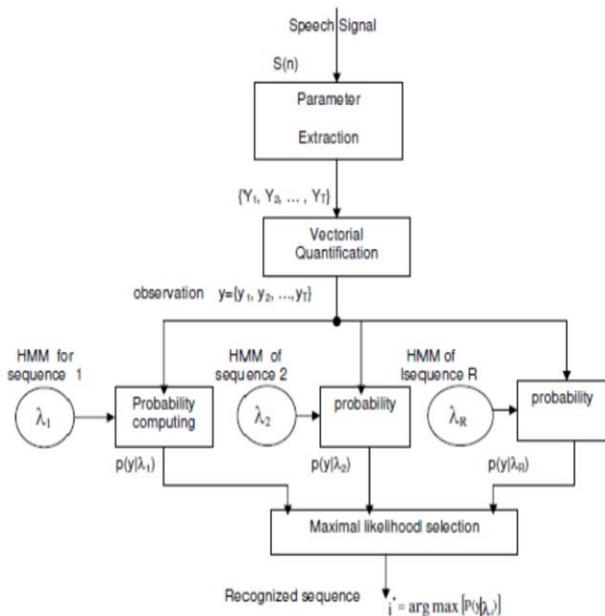


Fig. 5. HMM Algorithm.

Hidden Markov Models is a useful for data statistical modeling and classification.

The implementation of the HMM system requires three phases:

- Describe a network whose topology reflects the sentences, vocabulary words or basic units
- Make the training mode settings: $\lambda = (\pi, A, B)$
- Carry out the actual recognition occurrence by calculating the maximum likelihood [14].

IV. SIMULATION RESULTS

Several platforms and software are used in speech processing such as, Praat, Vocalab, EDVP, Speech Analyser Matlab and HTK toolkit. These tools offer many parameters and indicators form speech evaluation such as, pitch, formants, Jitter, shimmer and SNR.

A. Effect on the Pitch

Pitch is the first indicator of the speech production as represents the period of the glottis signal. Fig. 6 shows the variations of the speech waveform, the zero-crossing, the pitch, the spectrogram, the spectrum and the formants of a normal speaker (without any disease). We can observe that the pitch (Fig. 6) is characterized by a continuous and constant profile with of value $F1 = 210$ Hz (male speaker).

However, in the case of a pathological voice (organic or neurological origin), the speech profile presents distortions and dynamic variations around the pitch nominal value as illustrated in Fig. 7 to 10 which will be discussed later in details.

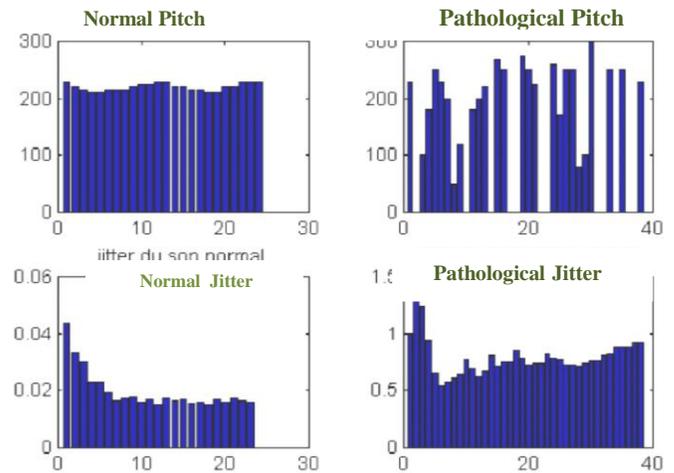


Fig. 6. Jitter Variation of Normal and Pathological Voice.

B. Effect on the Jitter and Shimmer

Disturbances of the durations of glottal cycles (Jitter) are irregularities in the period glottal signal. These disturbances are a basic phenomenon that is present in the voice and are therefore a feature of vocal timbre. This disturbance can be used to characterize spectrally hoarse voices, neurological, emotional or normal. The method is based on a study of the spectral effects of the glottal cycle's variance and the

evolution of the jitter values. On the other side, the study of perturbations of the amplitude (Shimmer) shows that they are a consequence of disturbances durations and energy. These mechanisms are asymmetries in the movement of both vocal cords and acoustic propagation of the glottal signal through the vocal tract [17]. Fig. 6 demonstrates that the normal jitter value is 0.02 (2%), hence the pathological value is over 0.8 (80%).

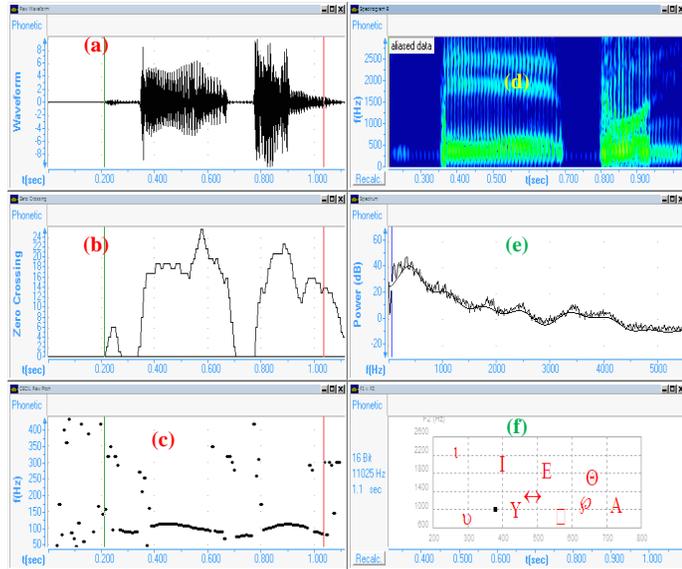


Fig. 7. (a) Speech, (b) Zero-Crossing, (c) Pitch, (d) Spectrogram, (e) Spectrum and (f) Formants Profiles (Case of a Normal Voice).

V. DISEASE IDENTIFICATION AND CLASSIFICATION

A. Multiple Sclerosis Disease

In this case of disease, the most common deficits affect recent memory, attention, processing speed, speech, visual-spatial abilities and executive function.

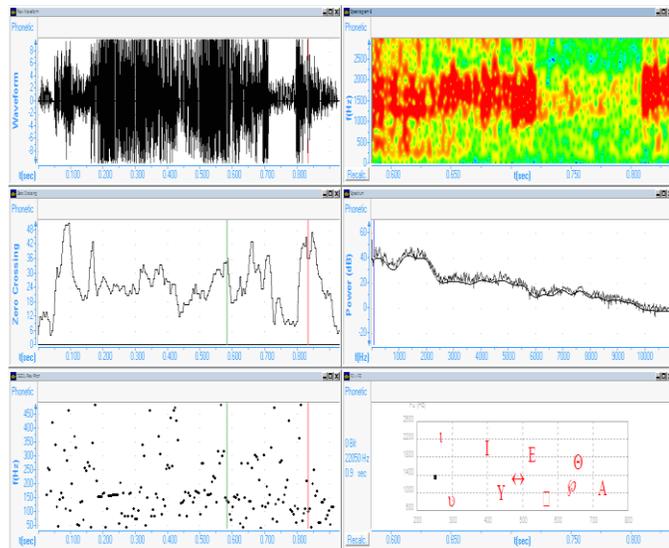


Fig. 8. Speech, Zero-Crossing, Pitch, Spectrogram, Spectra and Formants Profiles. (Case of pathological voice: Multiple sclerosis).

Symptoms related to cognition include emotion, instability and fatigue including neurological fatigue [18]. The following speech profiles and parameters of Fig. 8 (pitch, formants) demonstrated a correlation between the speech features and this pathology.

In fact, we can observe in Fig. 8 that in this case, the pitch profile becomes very disturbed with a high standard deviation contrarily with normal and safety speaker of Fig. 7. This state indicates a dysfunctioning of the speech production system which is monitored by the brain.

B. Dyslexia

It is considered a cognitive disorder but it does not affect intelligence. Problems may include difficulties in spelling words, reading quickly, writing words, "sounding out" words in the head [19], [20]. The examination of the speech of figure 9 shows disturbances of the pitch curve at the level of the co-articulations and the changes of vowel consonants in the pronounced word. These variations remain around 20% of the nominal value, but the standard deviation remains almost 12%. Also, the wide band and narrow band spectrograms are affected by this variation and disturbance.

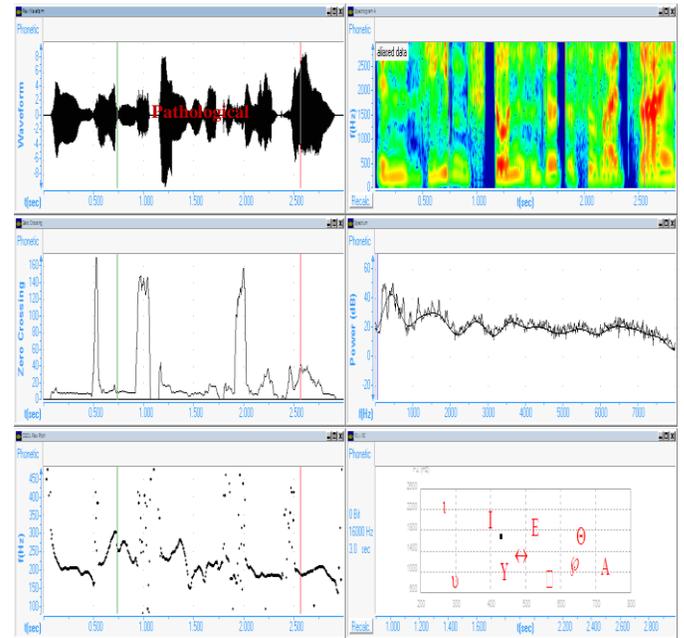


Fig. 9. Speech, Zero-Crossing, Pitch, Spectrogram, Spectra and Formants Profiles. (Case of pathology: Dyslexia).

C. Alzheimer

It is a chronic neurodegenerative disease that usually starts slowly and worsens over time. It is the cause of 60% to 70% of cases of dementia. The most common early symptom is short-term memory loss), problems with language, speech, disorientation, mood swings, loss of motivation and behavioral issues [21]. Recent research studies demonstrate relations between speech production and Alzheimer disease. Fig. 10 illustrates the speech parameters of a speaker suffering from Alzheimer (age: 72).

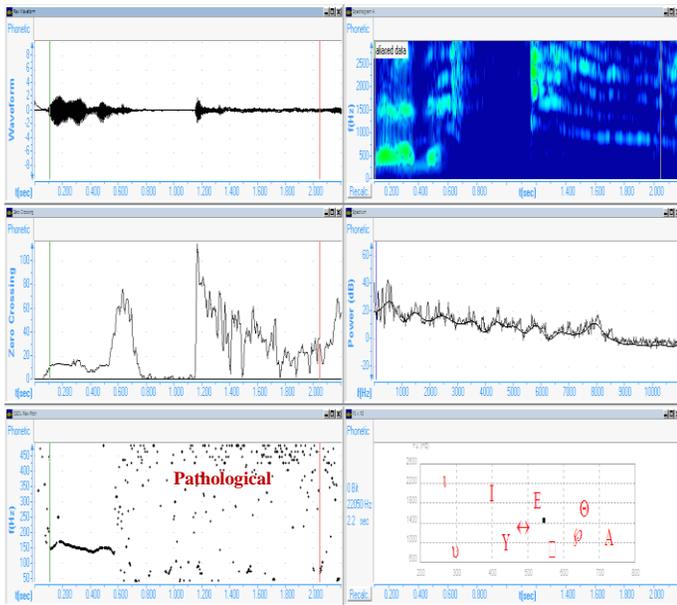


Fig. 10. Speech, Zero-Crossing, Pitch, Spectrogram, Spectra and Formants Profiles.

(Case of pathology: Alzheimer)

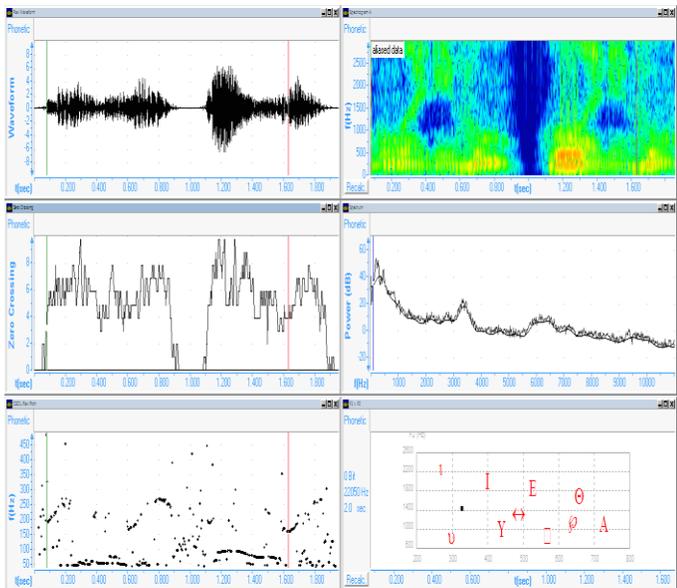


Fig. 11. Speech, Zero-Crossing, Pitch, Spectrogram, Spectra and Formants Profiles. (Case of pathology: Parkinson).

The examination of speech shows a loss of controls of the glottis after the pronunciation of the first word. No value of the pitch can be extracted due to the complete dispersion of the acquired values. The spectrogram of the same figure confirms these results and dot give us any information about the value of neither pitch nor the formants.

D. Parkinson

This neurology disease has a long-term degenerative disorder of the central nervous system that mainly affects the motor system. The symptoms generally come on slowly over time. At a first step, the most obvious are shaking, rigidity,

slowness of movement, and difficulty with walking. Thinking and behavioral problems may also occur. Dementia becomes common in the advanced stages of the disease. Depression and anxiety are also common occurring, including sensory, difficulty of speaking, sleep, and emotional problems [22]. The examination of the speech through Fig. 11 shows high deviations and variations of the pitch (glottis signal) rather large of 100% which alters the language understanding and loses the speech recognition of the speaker.

E. Pathology Recognition Ratio

All the described procedures and steps (acquisition, training, feature extraction, recognition and pathology classification) are embedded and inserted in a smart interface illustrated in Fig. 12.

Our tests are compared to the MEEI of KAY Pentax speech database, described in the last paragraph, on the HTK toolkit. According to Table II, we obtained a pathology recognition ratio (RR) between 86% and 100%.

These values are very interesting because they can discriminate several diseases from the normal voices characterized by a 100% value.

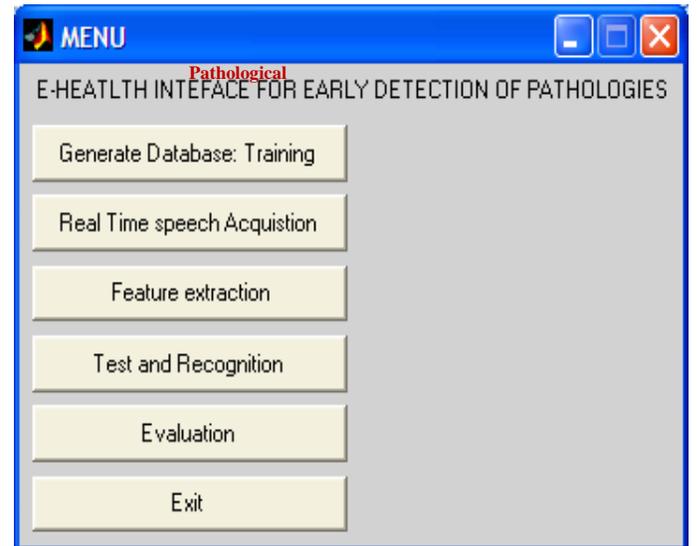


Fig. 12. Illustration of the E-Health Interface.

TABLE II. PATHOLOGY RECOGNITION RATIO OF SEVERAL SPEECH DISEASES

Disease	Recognition Ratio (RR) in %
Dysarthria	94
Gastric Refux	97
Ventricular compression	86
Nodule	92
Odeme	90
Disphonia	98
Normal	100

VI. DISCUSSION

The obtained Pathology identification ratios demonstrate that we obtain high (RR) values for both hyper-function (dysarthria) and paralysis diseases (dysphonia) respectively 94% and 98%. Besides, we compared our results with other studies using similar and different databases.

Table III compares our proposed algorithm with previous significant works [5, 6, 7, 8, 23, 24]. Although in these works the databases are different, it is observed that the proposed algorithm with HMM appears competitive and has a high accuracy to identify pathological and normal voices. We succeeded at the first step of our work to identify more than 8 kinds of organic and neurological diseases.

TABLE III. COMPARISON OF PERFORMANCES OF SEVERAL PATHOLOGY STUDIES

Reference	Database	Code book	Feature extraction	Classifier	Pathology identification
Godino	MEII	226	MFCC	SVM	94
Fonseca	MEII	154	MFCC	GMM	78
Souissi	Sharbrok	110	MFCC+D+DD	SVM	86
Boyanov	Private	400	Energy	LDA	93
Our method	MEII	657	MFCC+ Jitter + Shimmer	HMM	86 to 100

VII. CONCLUSION

In this paper, we developed a new tool dedicated to identification and diagnosis of vocal and neurological diseases. The method is based on analysis of acoustic parameters of a patient after a real time speech acquisition and processing. The modeling and classification procedures are automated by using HMM, training and recognition procedures. The validation was carried out thanks to the pathological famous database MEEI of Pentax. The obtained recognition ratio of the pathology is around 95%. The most significant indicators of the pathological speech are disturbances in amplitude (Shimmer) and frequency distortion and irregularities of pitch (Jitter) and finally the loss of glottis control (high standard value of the pitch). Besides, this application allows us to follow changes in the physiological state (heartbeat, blood pressure, ECG) and acoustic parameters (pitch, formants, timbre) and then we can compare them with normal and standard values. This is very interesting because it helps us to follow the disease evolution, to predict and to avoid patient complication and to improve his re-education therapy.

The following step of this study is to extend this application to other critical diseases such as cancer and Hepatics C and then to evaluate it through a large number of patients.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Biomedical Studies of Tunis. The authors declare no conflict of interest.

ABBREVIATIONS

HMM : Hidden Markov Models

GMM : Gaussian Mixture Model

MFCC : Mel Frequency Cepstral Coefficients

LPC : Linear Predictive Coding

RR: Pathology recognition rate:

HTK : HMM Toolkit

RASTA-PLP : Perception Linear coefficients

SVM : Support Vector Machines

LDA : Linear Discriminate Algorithm

REFERENCES

- [1] Maguire C., de Chazal P., Reilly R.B., and Lacy P. Automatic Classification of voice pathology using speech analysis. In *Proceedings of the World Congress on Biomedical Engineering and Medical Physics*, Sydney, 2003.
- [2] Boyanov, B. and Hadjitodorov, S. Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. *IEEE Engineering in Medicine & Biology Magazine* 16:74-82, 1997.
- [3] Manfredi, C., D'Aniello, M., Brusciagioni, P. and A. Ismaelli. A comparative analysis of fundamental frequency estimation methods with application to pathological voices. *Medical Engineering and Physics* 22:135-147, 2000.
- [4] Kay Elemetrics Corp. *Disordered Voice Database and Program*, Model 4337. Lincoln Park, NJ, Kay Elemetrics Corp. Canada 2010.
- [5] H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Voice pathologies identification speech signals, features and classifiers evaluation," in *Proc. Signal Process., Algorithms, Architect., Arrangements, Appl. (SPA)*, Sep. 2015, pp. 81–86.
- [6] F. Amara, M. Fezari, and H. Bourouba, "An improved GMM-SVM system based on distance metric for voice pathology detection," *Appl. Math*, vol. 10, no. 3, pp. 1061–1070, 2016.
- [7] J. I. Godino-Llorente, P. Gómez-Vilda, N. S-Lechón, M. Blanco-Velasco, F. Cruz-Roldán, and M. A. Ferrer-Ballester, "Support vector machines applied to the detection of voice disorders," in *Proc. Int. Conf. Nonlinear Anal. Algorithms Speech Process.*, vol. 3817. 2005, pp. 219–230.
- [8] N. Souissi, C. Adnen, "Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine," in *Proc. 7th Int. Conf. Modelling, Identificat. Control (ICMIC)*, Dec. 2015, pp. 1–6.
- [9] I. M. M. El Emary, M. Fezari, and F. Amara, "Towards developing a voice pathologies detection system," *J. Commun. Technol. Electron.*, vol. 59, no. 11, pp. 1280–1288, 2014.
- [10] L. Verde, G.D. Pietro, G. Sannino. *Voice Disorder Identification by Using Machine Learning Techniques*. *IEEE Access journal*, Vol 6, April 2018. DOI: 10.1109/ACCESS.2018.2816338
- [11] NINDS Multiple Sclerosis Information Page". *National Institute of Neurological Disorders and Stroke*. November 2015.
- [12] Nakahara J, Maeda M, Aiso S, Suzuki N. "Current concepts in multiple sclerosis: autoimmunity versus oligodendroglialopathy." *Clinical reviews in allergy & immunology*. 42 (1):26-34. 2012. DOI :10.1007/s12016-011-8287-6.
- [13] J. Wang and C. Jo, "Vocal folds disorder detection using pattern recognition methods," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS)*, Aug. 2007, pp. 3253–3256.
- [14] R. T. Ritchings, M. McGillion, and C. J. Moore, "Pathological voice quality assessment using artificial neural networks," *Med. Eng. Phys.*, vol. 24, nos. 7–8, pp. 561–564, Sep. 2002.
- [15] Young S. J., Woodland P. C., Byrne W. J., "HTK Reference Manual for HTK Version 3.1", 2001.

- [16] L. Bailly. Interaction between the vocal cords and ventricular strips in phonation: exploration in vivo, physical modeling, validation in-vitro. Doctoral thesis, University of Nantes, France, 2009.
- [17] B. Teston. The objective assessment of dysphonia: current state and prospects of evolution.. Report of the Laboratory speech and language, University of Provence, CNRS France, 2011.
- [18] Bobholz J, Rao S. "Cognitive dysfunction in multiple sclerosis: a review of recent developments". *Current Opinion in Neurology*. **16** (3), 2003. doi:10.1097/00019052-200306000-00006.
- [19] Phillips, Sylvia; Kelly, Kathleen; Symes, Liz . *Assessment of Learners with Dyslexic-Type Difficulties*. SAGE. p. 7, 2013. ISBN 978-1-4462-8704-0.
- [20] Campbell, Robert Jean. *Campbell's Psychiatric Dictionary*. Oxford University Press, 2009, pp.310–312. ISBN 978-0-19-534159-1.
- [21] Querfurth HW, LaFerla FM (28). "Alzheimer's disease". *The New England Journal of Medicine* **362** (4):329–44. January 2010. doi:10.1056/NEJMra0909142.
- [22] Sveinbjornsdottir, S. "The clinical symptoms of Parkinson's disease.". *Journal of Neurochemistry*. **139**: 318–324. July 2016. doi:10.1111/jnc.13691.
- [23] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases," *IEEE Eng. Med. Biol. Mag.*, vol. 16, no. 4, pp. 74–82, Jul. 1997.
- [24] S. Jothilakshmi, "Automatic system to detect the type of voice pathology," *Appl. Soft Comput.*, vol. 21, pp. 244–249, Aug. 2014.

An Overview of Mutation Strategies in Bat Algorithm

Waqas Haider Bangyal¹
Member IEEE SMC
Department of Computer Science,
Iqra University, Islamabad, Pakistan

Jamil Ahmad²
Senior Member IEEE, Professor Computer Science
Department of Computer Science,
Kohat University of Science and Technology (KUST),
Kohat, Pakistan

Hafiz Tayyab Rauf³
Department of Computer Science,
University of Gujrat,
Gujrat, Pakistan

Sobia Pervaiz⁴
Department of Software Engineering,
University of Gujrat,
Gujrat, Pakistan

Abstract—Bat algorithm (BA) is a population based stochastic search technique encouraged from the intrinsic manner of bee swarm seeking for their food source. BA has been mostly used to resolve diverse kind of optimization problems and one of major issue faced by BA is frequently captured in local optima meanwhile handling the complex real world problems. Many authors improved the standard BA with different mutation strategies but an exhausted comprehensive overview about mutation strategies is still lacking. This paper aims to furnish a concise and comprehensive study of problems and challenges that prevent the performance of BA. It has been tried to provide guidelines for the researchers who are active in the area of BA and its mutation strategies. The objective of this study is divided in two sections: primarily to display the improvement of BA with mutation strategies that may enhance the performance of standard BA up to great extent and secondly, to motivate the researchers and developers for using BA to solve the complex real world problems. This study presents a comprehensive survey of the various BA algorithms based on mutation strategies. It is anticipated that this survey would be helpful to study the BA algorithm in detail for the researcher.

Keywords—Bat algorithm; optimization; local optima; mutation strategies; premature convergence; swarm intelligence

I. INTRODUCTION

From last two decades, optimization [1] has been considered as the most active area of research. Advanced optimization algorithms are required as the real world's optimization problems shifting towards complexity. The major goal is to reach the optimal value of the fitness function. In real-life applications, optimization is mainly incorporated for nonlinear complex problems. It provides optimum solution by searching a vector in function. Available solutions are incorporated from possible values, however, optimum solution is the intense value [2]. The goals of optimization are to decrease cost, wastes and time, or increase performance [3], profits and benefits. This is the fact that established deterministic procedures or algorithms do not resolve a huge amount of data problems in real life.

Swarm Intelligence (SI) belongs to artificial intelligence domain, used to design multi-agent intelligent systems by

embedding the qualities of social insects' behavior in the form of swarms such as bees, wasps and ants, and flocks of fishes and birds. SI first inaugurated by Beni [4] in cellular robotics system. Researchers are attracted with the behavior of flocks of social insects from many years. The individual member of these flocks can achieve hard tasks with the help of collaboration among others [5]; however, such colonies behave without any centralized control. SI is very famous in bio-inspired algorithms, computer science, and computational intelligence. In addition, this SI nature inspired Meta Heuristics algorithms [6] are commonly implemented for solving optimization problems as well as computational intelligence. As compared to traditional algorithms, particle swarm optimization, ant and bee algorithms, firefly and cuckoo search algorithms that based on SI are more beneficial.

A meta heuristic algorithm of SI family named Bat Algorithm (BA) introduced by Xin-She Yang, which works on the micro-bats' behavior of echolocation [2], with variation of pulse rate loudness and emission. The objective of BA is that they fly randomly with the variation in their velocity and frequency and find out their prey in search space [7]. Echolocation is the most essential feature of bat behavior, which means that bat produces a sound pulse and listen the echo that is bouncing back with collision of obstacle during flying [8]. BA is a novel approach based on the two important parameters: pulse rate and loudness [9], and both parameters are fixed while BA is executed.

For researchers, the most important issue is swarm convergence; they are still confused about the convergence of bat towards same curve. They are assumed to highlight the parameters, which causes swarm convergence. Bats' premature convergence [10] greatly affects the performance of BAT algorithm. Before global optimum solution is found, swarms stuck in local optima because of premature convergence. To resolve local optima issues, researchers have introduced many improve methods [11], where divergent mutations are executed on applicable parameters like frequency, pulse rate, velocity, swarm size and loudness [12].

Two most essential parts of meta-heuristic BA that needs to be balanced [7], are exploration and exploitation. Exploration

is used to find out the global optimum solution while the exploitation is used to capture the local minimum solution. Without trapping into local optima, BA must have to achieve the global optimum solution. Although, the balance between both the components could cause the success of BA. A small amount of exploration [8] and a large amount of exploitation generates premature convergence, while a large amount of exploration and the small amount of exploitation have become the reason to create difficulties for algorithm towards the convergence in local optimum.

Premature convergence is the biggest issue of BA due to the exploration and exploitation [9]. However, researchers are still not sure about the convergence of bat and supposed [10] to point out all the parameters which cause the convergence. The researchers introduced various technique and BA variants to overcome the premature convergence in BA with changing the features like size, velocity, pulse emission rate, loudness, and frequency [6]. However, they succeeded to reduce the effect of premature convergence in BA with their enhanced BA version [11].

The rest of the paper is organized like this the standard BA is discussed in Section II. A comprehensive over view on BA mutation strategies is presented in Section III. Discussion is encompasses in Section IV and future direction for further enhancement is elaborated in Section V.

II. BAT ALGORITHM

The working of BA is quite similar to the natural behavior of bats. They use their echo to search the entire food; similarly, BA adopted the echolocation of micro bats in order to find the global best solution. BA follows the three basic rules: The first one is estimating the optimal distance to the food using the phenomena of echolocation. Secondly, population moves into the search space with distinct velocity and fixed frequency. However, the wavelength and bat loudness can vary according to their distance between food and the entire bat current position. The third rule followed by BA is linearly decreasing behavior of bat loudness factor.

BA includes a candidate solution which is revealed by the bat population, for $i = 1 \dots N_p$ the candidate solution is expressed by the solution vector $x_{ik}^t = (x_1, \dots, x_{ik})$ along each dimension d within real value components x_{ij}^t , where the vector interval for each real value component is $x_{ij}^t \in [x_l - x_u]$ and N_p representing the size of population. While, x_l and x_u defines the upper and lower limits of current vector solution [9]. The principal ingredients of BA are population initialization, mutation procedure, exploitation, exportation, and updating of the current best solution.

Step 1: In this phase of population initialization, the bats assigned an initial value considering as their current best vector position and velocity. Those initial values can be generated using uniform distribution at random locations.

Step 2: This phase includes the echolocation of bats, in which the mutation operator is carried out to simulate and designates the implicit individuals (bats) into the entire search

space in order to find the temporary initial solution. The frequency of bats is represented by the following equation.

$$f_i = f_{min} + (f_{max} - f_{min}) \cdot R(0,1) \quad (1)$$

f_{max} and f_{min} are the maximum and minimum threshold of the bats frequency which is vary according to the nature of problem where $R(0,1)$ is a random number following the uniform distribution.

Bats used the following equation in order to update their current velocity v_{ik}^{t+1} :

$$v_{ik}^{t+1} = v_{ik}^t + (x_{ik}^t - p_i^t) f_i \quad (2)$$

At iteration, v_{ik}^t denotes to bats old velocity and p_i^t is current global optimum. x_{ik}^t is referred as the current bat position.

The updated bat position x_{ik}^{t+1} can be calculated using the following equation:

$$x_{ik}^{t+1} = x_{ik}^t + v_{ik}^{t+1} \quad (3)$$

Step 3: In this phase of bat exploitation, a random walk is employed to adjust the current best solution. eq.4 is used to generate new solutions

$$x_{new}^t = p_i^t + \varepsilon A_{ik}^t \quad (4)$$

A_{ik}^t is bat loudness which is decrease linearly against each iteration t where ε is a scaling factor generated randomly of interval $[-1,1]$.

Step 4: The pulse rate r_{ik}^t of bats associated with the probability of the rate of omitting pulse in the particular iteration t . This probability highly depends on the loudness of bats A_{ik}^t . Basically, the pulse rate r_{ik}^t and the loudness A_{ik}^t of bats are two major controlled parameter employed to control the convergence rate of BA. Commonly, the loudness of bats A_{ik}^t tends to decreases and the pulse rate r_{ik}^t tends to increases when the bats nearly close to their best solution (Fig. 1 and 2). The loudness of bats raises and pulse rate declines when the bat gets its prey, so these both characteristics mimic the original bats.

Both control parameters are denoted by the following equations:

$$A_{ik}^{t+1} = \alpha A_{ik}^t \quad (5)$$

$$r_{ik}^t = r_{ik}^0 [1 - \exp(-\gamma^t)] \quad (6)$$

In the eq.5, α representing an constant usually fixed according to problem nature, this parameter is used to control the convergence of bats. Where eq.6 contains constant number as γ . Flow chart of Standard Bat Algorithm is given in Fig. 3. The pseudo code for standard BA is presented in Algorithm 1.

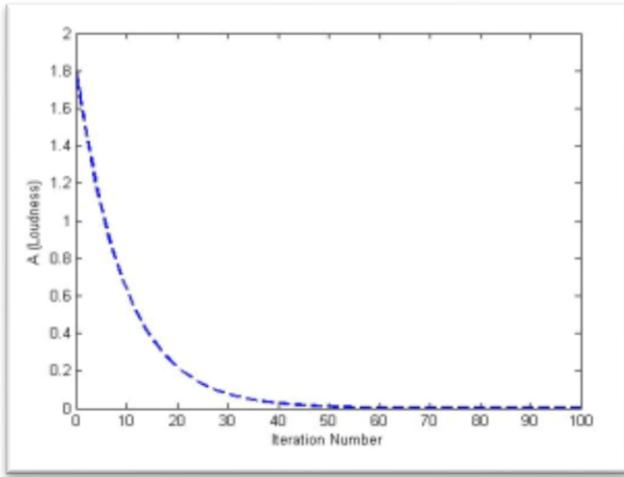


Fig. 1. Decreasing in loudness

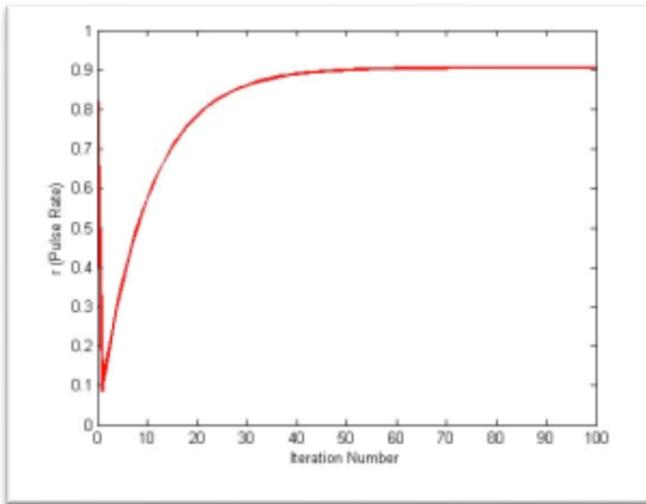


Fig. 2. Increasing in pulse rate

- (15) $x_{ik}^t \leftarrow x_{new}^t; f_{old}^{min} \leftarrow f_{new}^{min};$
- (16) **end if**
- (17) $f_{new}^{min} \leftarrow \min(f(x_{ik}^{best}));$
- (18) **end for**
- (19) **end while**

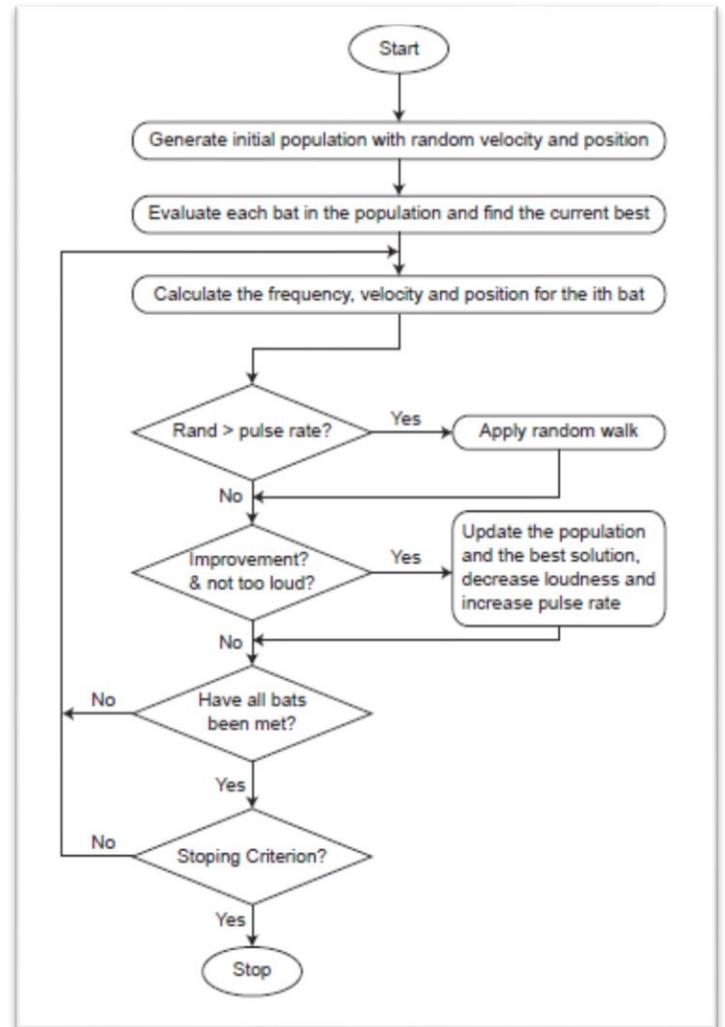


Fig. 3. Flow chart of Standard Bat Algorithm

Algorithm 1: Standard Bat Algorithm

Input: $x_{ik}^t = (x_1, \dots, x_{ik})$

Output: x_{ik}^{best} & $\min(f(x)) \rightarrow$ global optimum and minimal fitness function score

- (1) $x_{ik}^t \leftarrow \text{Init}(x_i)$; initialization at random location
- (2) $\text{Initial_fitness}(x_{ik}^t)$;
- (3) $\text{Calculate}(x_{ik}^{best})$; Current global optimum
- (4) **While** ($t \leq t_{max}$) **do**
- (5) **for** $i \leftarrow 1 \dots N_p$ **do**
- (6) Update f_i by eq.1
- (7) Update v_{ik}^{t+1} by eq.2
- (8) Update x_{ik}^{t+1} by eq.3
- (9) **if** $r(0,1) > r_{ik}^t$ **then**
- (10) $x_{new}^t \leftarrow$ update current solution by eq.4
- (11) **end if**
- (12) $f_{new} \leftarrow$ Compute x_{new}^t
- (13) $\text{eva} \leftarrow \text{eva} + 1$;
- (14) **if** $f_{new}^{min} < f_{old}^{min}$ **and** $\text{Ra}(0,1) < A_{ik}^t$ **then**

III. MUTATION STRATEGIES

Swarm convergence has been prominent issue for the researchers till now researchers are confused to decide whether particles converge to the same curve or not. They are supposed to highlight [13] the parameter that plays major role in swarm convergence [14]. A mutated BA helps the particles to avoid premature convergence around local optima [15].

Table I shows the issued paper regarding mutation strategies in the highlighted time frame. It depicts the larger number of researchers shown the interest in this field. Multiple components impact this increment from 2011 to 2018 in publication. Moreover, it shows the awareness and significance of mutation strategies in different areas. Interestingly, Table I shows the maximum number of articles published in journals as opposed to the conference.

TABLE I. YEAR WISE STATISTICS WITH DETAILED OBSERVATION

Year	2011	2012	2013	2014	2015	2016	2017	2018	Total
Journal	2	3	4	11	7	7	2	2	38
Conference	-	-	2	3	8	8	4	-	25
Total	2	3	6	14	15	15	6	2	63

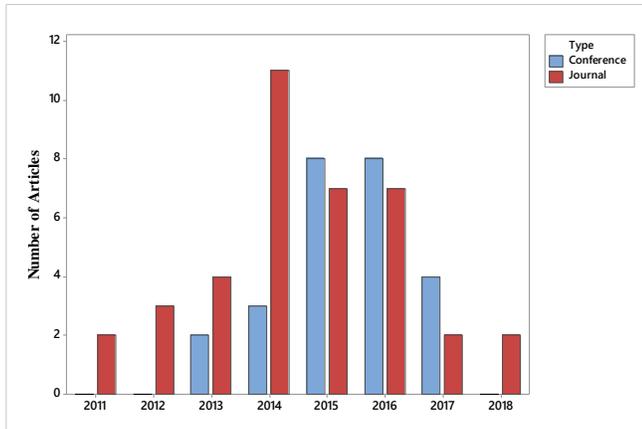


Fig. 4. Year wise representation of number articles

The objective of this paper is to give a review of related work, which is performed on Meta heuristic algorithms. The author in [16], introduced a new algorithm named as Generalized Evolutionary Walk Algorithm (GEWA). In last three years, various improvements and hybridizations on Meta heuristics are introduced. In this paper, the author examines the main elements of such algorithms and describes their work flow.

For the solution of engineering optimization issues in [17], a new BAT is introduced, where the validity of new BAT has been measured with consideration of seven benchmark problems of engineering design. The results illustrated that new BAT is significant as compared to others.

BAT is famous to resolve the problems of nonlinear global optimization, thus in [6] a solution of multi-objective optimization has been given, named Multi-objective Bat Algorithm (MOBA). Firstly, proposed approach validation is performed on standard sub-set of test function and after that against the design problems related to multi-objective like welded beam design. Exhaustive analysis shows the efficiency of proposed approach.

In order to solve binary problems, in [18], the author introduced a new approach Binary Bat Algorithm (BBA). A comparative study is performed on twenty-two standard functions along with Binary Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to conclude the outcomes of new introduced approach. As well as, Wilcoxon's rank-sum nonparametric test is used to judge the BBA difference of results as compared to other algorithms. According to results, it is concluded that BBA gives outstanding results on most standard functions.

The author in [19], proposed a new modified technique Opposition-Based BAT algorithm (OBA). To improve the global optimal solution and convergence speed, opposition

based methodology is integrated with standard BAT's foundation. According to the comparison results, conducted with other algorithms like real coded genetic algorithm (RGA), PSO and DE, effectiveness of proposed algorithm is proved in terms of efficiency.

Literature of last three year described in [20], the goal of this paper is to explain all working of BAT and its new versions as well as its applications. Moreover, the author of this paper precisely describes the case studies. In conclusion, future work related to BAT is also mentioned.

In [21], author proposed an Improved Bat Algorithm (IBA) for solving the insufficient performance of traditional BAT Algorithm. The author explored the exploitation and exploration strategies with three modifications.

A new hybrid method introduced in [22], which combines BAT with the help of mutation operator of differential evolution and beta probability distribution (BADEBD). This method is proposed to solve the problems related to Jiles-Atherton (J-A) modeling. Test results produced significant outcomes as compared to simple BAT Algorithm.

To improve the performance, the author of [23] proposed a modification in hybridized BAT algorithm called a modified BAT-inspired differential algorithm. Modification is performed on mutation and crossover as well as for the stability of local and global search, a novel pulse rate function and loudness introduced. The analysis has performed on five standard test function, which concludes that proposed algorithm give outstanding results.

The author of [24] provided details of new technique BAT Algorithm with Gaussian Walk (BAGW), which is working to overcome the problems related to high dimensions in search spaces. In the paper, the operators: local search, velocity and frequency are helping to achieve BAGW for higher computational results. In the paper, analysis is performed on 4 benchmark functions, which provide the excellence of BAGW as compared to standard BAT.

The author in [25], proposed a new approach Chaotic Bat Algorithm, this approach is implemented by introducing chaos in BAT. The purpose of this introduced approach is to provide stable and strong global optimization by increasing the mobility of its global search. The comparative results of these variants shows that some chaotic variations outperform the benchmark BAT for standard functions.

A critical analysis of Swarm Intelligence (SI) algorithms explained [26], in which the behavior of evolutionary operators is examined. Furthermore, the working of crossover, mutation and selection operators is investigated to achieve exploitation and exploration. Beside this, algorithms are tested on self-organization, Markov chain framework and dynamic system. At last, future work is also discussed.

To overcome the limitation in population diversity, [27] presented a novel complex-valued bat algorithm. The introduced approach implemented the complex-valued encoding, so the imaginary and real parts will be individually updated. Beside this, introduced approach significantly expands the diversity towards the population as well as

dimensions are also expanding for indication. Comparison has been performed with PSO and fourteen test functions, which produced feasible results.

To figure out real world complex problems, an improved method of bat algorithm with chaos initiated. The method has been implemented for permutation of random number generation (RNG) with the use of chaotic sequences to initialize parameters [28]. Standard test functions used to investigate the validity of Chaotic Bat algorithm (CBA), which demonstrate efficiency of CBA than traditional BAT.

Towards the solution of binary space's optimization problem, S. Sabba et al. [29] inaugurated discrete binary bat algorithm (BinBA). The sigmoid function is embedded in proposed algorithm, which implemented by Kennedy in 1997 for binary PSO. The results are promising, when BinBA compared with multidimensional knapsack problems.

BAT is not good in explorative features, to sort out this problem a Novel Adaptive Bat Algorithm (NABA) introduced by Kabir et al. [30]. Two strategies embedded to enhance BAT exploration's degree: Gaussian probability distribution for production of step size in mutation and Rechenberg's 1/5 mutation rule. These strategies embedded to balance exploration and exploitation. Simulation results of NABA showed its compatibility, when conducted with test functions and comparison performed with traditional BATS.

In continuous domain, to resolve global numerical optimization, a self-adaptive bat algorithm (BA-SAM) is presented by Kabir et al. [31] where two enhanced search equations are introduced. Moreover, selection probability used to handle the frequency to apply the introduced equations that results in new BA-SAM. Experimental results showed that BA-Sam outperforms than traditional BAT and other algorithm, when tested on both uni-model and multi-model continuous test functions.

To remodel the echolocation model of BAT with proper utilization of cloud model, a new approach named Cloud Model Bat Algorithm (CBA) introduced in [14], to qualitatively represents the concept "Bats approach their prey". Moreover, for balancing of exploration and exploitation, population information communication and levy flight mode proposed. Experimental results are taken from function optimization, described that CBA is good in performance

For reduction of premature convergence in Bat swarm optimization (BSO), Chaotic based technique incorporated in BSO [32], to introduce Chaotic Bat Swarm Optimisation (CBSO). A best chaotic technique selected from available eleven chaotic map functions for CBSO. CBSO analysis performed on various test functions, which demonstrated that CBSO is good in quality than other algorithms. To represent the solution of weights and ANN architecture optimization, in [33], a novel modified bat algorithm proposed. For the improvement of BAT exploration and exploitation abilities, two modifications introduced. Throughout the training process, various versions of introduced BAT incorporated to figure out selection of architecture, weights and ANN's biases. The test was performed on six classifications and two datasets of standard time series, which showed satisfactory results.

To solve optimization problems related to discrete and continuous data by implementing BAT. Thus, experimental studies proved that the main problem is to set control parameters, and it consumes maximum time to provide its best combination of parameter. The paper is proposed a new advancement in BAT, which implemented without control parameters. Initially in [34], this version is tested on standard benchmark functions that shows the capability of this advanced version.

A Multi-Objective Binary Bat Algorithm (MBBA) proposed in [35], to handle binary search space problems associated with multi-objective optimization. The algorithm implemented an advanced BAT position upgrading method that performs effectively with binary search space. For enhancement in local search capabilities a mutation operator is initiated, to find heuristic Pareto Solution a Pareto dominance approach is used, and for BATs' flight a flight leader selection method. The MBBA is tested with non-dominated sorting genetic algorithm 2 (NSGA-2), and results proved MBBA performance is better.

To resolve the problem of global search, in [15], a Multi-Swarm Bat Algorithm (MBA) is introduced. With the help of parameter settings, information is transferred among various swarms through immigration operator. Although, this structure can produce good balance between local and global search. Furthermore, in swarm's elite, a best individual is passed by using selector operator. The BAT individual transferred towards next generation without using any operator, which ensures that throughout optimization process, these solutions cannot be damage. To evaluate the effectiveness of MBA, it is compared with actual BAT on sixteen standard functions, which conclude that MBA produce more significant values of functions than BAT.

For the solution of exploitation abilities in [36], authors introduced a new parameter called inertia weight that enhance the performance of exploitation ability. The parameter is used to control the effect of inertia of all BATs prior velocity. To test the outcomes, CEC2013 standard problems are verified through it, and experimental result shows the author's method validity.

For target matching, a novel approach introduced in [37], named as Chaotic Mutated Bat Algorithm (CMBA) Optimized Edge Potential Function (EPF). The proposed technique is suitable for the correctness and stability, used for target recognition and CMBA is used to optimize matching parameters. Real-world applications are experimented for the verification of introduced approach feasibility. The simulated results exhibited that introduced approach works effectively throughout the target matching.

In [38], author described a BAT Algorithm improvement by using fuzzy system that adapt its parameters, dynamically. The described method is compared with standard BAT and also with GA, which gives a more effective analysis related to BAT Algorithm. The exhaustive analysis proved that the proposed method produces outstanding results than original BAT Algorithm and Genetic Algorithm.

A new modified version of BAT Algorithm (BAT) introduced in [39], called as Enhanced BAT algorithm (EBA). Local and global searching features of BAT are modified to introduce a new version of BAT by using three different methods. EBA technique is tested on different standard functions that give excellent results as compared to BAT.

A new approach of BAT proposed called novel Bat algorithm (NBA) in [12] that deals with the behavior of BATs. For designing a novel local search technique, the author embedded environment selection of BATs and their self-adaptive remuneration in standard BAT for Doppler effect in echoes. Analysis is performed on four real life engineering problems and twenty standard functions, which demonstrate the superiority of NBA as compared to other algorithms.

To deal with constrained optimization a technique Novel Hybrid Bat Algorithm with Differential Evolution strategy (BADE) introduced in [40]. With the consolidation of DE with BAT, the intrusion of BATs can be significantly imitated by BADE. The impact of other BATs can be calculated by changing the velocity equation of BAT through the addition of mean velocity of swarm. The verification of introduced algorithm is done with nine standard functions and three problems related to engineering designs, which proved that the introduced algorithm is more effective.

In [3] a new variant of BA name Accelerated Bat Algorithm (ABATA) is proposed to improve the local search ability using the Nelder-Mead strategy. The proposed ABATA is tested on seven problems based on integer programming method. The results demonstrate that the proposed method performs better as compared to the classical methods.

The authors in [41], introduced a new variant of BAT algorithm with name Hybrid BAT. Additionally, some techniques of population initialization, decoding and encoding are also implemented to modify BAT for PP problems. Two techniques of local search are implemented in BAT to reduce local convergence. For the diversity of BATs' community, two novel operators for solution representation are also proposed. The experimental results of hybrid BAT impressively shows the efficiency.

In [42] a modified Bat Algorithm for the Quadratic Assignment Problem: Quadratic Assignment problem (QAP) is basically related to discrete search space. This paper is introduced an approach, which handled QAP with respect to Bat Algorithm (BA). As BA is suitable for continuous problems, it cannot be implemented directly to resolve QAP. Therefore, Smallest Position Value (SPV) rule is used to apply BA for QAP, which gives the solutions with appropriate findings concerned with sequencing problems. According to statistical results, the introduced approach provides outstanding results in all cases of PSO. However, the most crucial output result is when input size is small then BA meets optimal fitness whether problem size is huge.

To resolve the problem of global optimization, in [43] author proposed a novel methodology called Variable Neighborhood Bat Algorithm (VNBA), in which Variable Neighborhood Search (VNS) is implemented with simple BAT algorithm in terms of local search. An exhaustive analysis is

performed on sixteen standard Test functions, which proved the outstanding performance of VNBA.

The author introduced a new methodology in [44] of BAT Algorithm, in which strategy adopts a dynamic behavior of BAT parameters. The author is used an Interval Type-2 Fuzzy Logic to implement the introduced strategy. The defined strategy is compared with Type-1 Fuzzy Logic, which shows the results in favor of Interval Type-2 Fuzzy Logic.

A new method Enhanced Shuffled BAT Algorithm (EShBAT) described in [45], which is an improved version of Shuffled BAT Algorithm (ShBAT). This method is implemented to increase the exploitation abilities of ShBAT where for the formation of super-memplex, it collects best memplex from many and form their groups. The super-memplex forms independently, to further utilize best solutions. The EShBAT is verified on 30 standard functions. As compared to BAT and ShBAT, EShBAT gives superior results.

The authors of [46] introduced a new algorithm with name BAT-PSO to handle the image registration issues, to significantly utilize in health care and research. According to statistical results, BAT-PSO does better search for optimal parameters registration.

To resolve the issues related to divergence of BAT and convergence speed, a new advanced BAT introduced in [47] named as Local Enhanced Catfish Bat Algorithm (LECBAT). To improve the exploitation capabilities, it keeps and uses the contradictions of first best and second best optimum search solutions. For the improvement of search precision, a dynamic adjustment is performed in evolution process on local scale element. If an algorithm produces similar global best for continuous iterations, then it is considered as local optimum. The results concluded with the comparison of various modifications in BAT shows the superiority of LECBAT.

In [48] Multi-Objective Reactive Power Dispatch in Distribution Networks using Modified Bat Algorithm: the author produced a new variant of BAT algorithm with name Modified Bat Algorithm to support voltage in distributed architecture.

To support voltage in distributed architecture this paper is introduced a new Modified BATt Algorithm for Optimal Reactive Power Dispatch (OPRF). The major concern of objective functions of introduced approach is to reduce reactive power wastage as well as prevent voltage infringements. An introduced approach has been tested by using a test node feeder of IEEE 37. The tool for simulation of algorithm is python; similarly, PowerFactory is used for architecture. Test results explained that the introduced approach is able to reduce wastage whereas also supply voltage regulation.

In [49] Review of Recent Load Balancing Techniques in Cloud Computing and BAT Variants Is proposed. The purpose of this paper is to study all research algorithms of Load Balancing, which is commonly used in current situation of Cloud Computing Environment. The most appropriate algorithms that are applicable in every emergent field, mainly inherited from nature. So, the paper is BATsed on one of the nature inspired technology called BATt Algorithm. The paper

is described a work, after comparing all recent BAT and Load BATlancing approaches.

In the paper, traditional BAT is hybridized with proposed method using tabu search, the selection procedure of new solution in traditional BAT is transformed. The target of proposed method is to increase the consistency of objective functions. The Hybrid BAT algorithm proved its superiority, when it is compared with other two standard algorithms in [50].

In [51] Modified Bat Algorithm for Localization of Wireless Sensor Network is introduced. The paper is proposed a technique for wireless sensor networks, which is used to estimate the node localization issues. In the meantime, with the use of bacterial foraging optimization algorithm's technique called bacterial foraging, the modification in traditional BA is performed. The simulation results of proposed method describe that its robustness is increased as well as it provides appropriate success in increment of localization ratio and a good convergence speed.

For community detection, four novel variants of BAT proposed in [52] by combining modularity metric and Hamiltonian function with enhanced discrete BAT and Novel BAT algorithm. Comparison of new variants had been performed with previous approaches like fast greedy modularity optimization, Grivan and Newman, Reichardt and Bornholdt, Ronhovde and Nussinoy and spectral clustering, which illustrated promising results of new variants.

The author in [53] described about the emergent field of nature-inspired meta heuristic BAT algorithm and its application, which showed the effectiveness and efficiency of BAT.

To figure out higher dimensional problems in search space and reduce the step size along with the use of BAT, N. M. Nawi et al. [10] proposed an approach named Improve Bat Algorithm with Gaussian distribution random walk (BAGD). The objective of proposed BAGD is to take short step during the period of search. Comparative analysis was performed with six metaheuristic algorithm based on ten standard test functions, which indicates BAGD performance is better than others.

A new algorithm named Bat Flower Pollination (BFP) introduced, to handle the synthesis of diversely separated linear antenna array (LAA) in [54]. The introduced approach works with both BAT and flower pollination algorithm (FPA). To avoid local minima, both algorithms collaborate with each other as well as also suitable for synthesis of diversely separated LAA. A validity of introduced BFP was confirmed through ten standard test functions along with other famous techniques. The results of test concluded that proposed BFP gives superior results than others.

BAT performance is not good for multi-model numerical problems as described in [55], so the author proposed optimal forage strategy. The strategy implemented to direct each bat for search direction. Moreover, random disturbance strategy also used to increase the pattern of global search. CEC2013 test suite was considered for verification of modified algorithm

along with four evolutionary algorithms. The results explained the effectiveness of modified approach.

To find global optimal solution with proper consistency and increase the exploration ability of BAT, D. Singh et al. [56] proposed modified bat algorithm (MBA). Experimental analysis showed MBA is better than others, when it was tested against standard benchmark problems as well as with stateOf0art algorithms.

To solve the BAT Algorithm, local search problem, a new method Velocity Adaptive Shuffled Frog Leaping BAT Algorithm (VASFLBA) is proposed by the authors in [57]. By the usage of meme transfer process with respect to Shuffled Frog Leaping Algorithm (SFLA), a local search capability is extended. Beside this, for improvement of global search capability, random population competition is proposed. A differential mutation process is implemented to advance the global search optimum, so the algorithm can stand out in local optimal, when an algorithm is trapped in local optimal. The validity of VASFLBA is tested with standard functions. Because of these experimental results, this algorithm is implemented for industrial control system (ICS), to optimize its parameters that belong to support vector machine (SVM). The experimental results show that VASFLBA is better in performance as compared to BAT, SFLA and other algorithms.

The authors introduced a new variant of BAT algorithm in [58] with name Improved Binary Bat Algorithm (IBBA). An IBBA is used to enhance the global search of BAT with the help of crossover, mutation and selection operators belong to Differential Evolution (DE). The experiment conducted with respect to PSO and GA, showed that IBBA is better in correctness and stability.

A novel hybrid algorithm (MDBAT) with hybridization of traditional BAT and multi-directional search algorithm (MDS) introduced by the authors in [59]. It is used to handle the problem belongs to global optimization. The author's goal is to reduce the slow convergence that is the reason, MDS is used for hybridization. MDS speed up the working of introduced algorithm rather than continues the iterations of BAT for a long time without any satisfactory results. The introduced algorithm is tested against sixteen global optimization issues as well as with eight standard algorithms. In conclusion, MDBAT outperforms than other algorithms.

A novel resolution method belongs to directional bat algorithm (dBA) is introduced, to handle reliability based design optimization (RBDO) by A. Chakri et al. in [60]. In the meantime, ϵ -constrained technique was also embedded into dBA, to resolve constrained optimization problem in effective manner. Various engineering problems were used for testing and results concluded that proposed method can solve RBDO problems.

An improved version of BA called I-BAT is proposed [61] to enhance the exploitation capabilities of BA. The authors used Torus distribution to initialize the bats in I-BAT and the controlled factor equal 0.1 is carried out for managing the search manner into the local area. Experimental Results proved I-BAT as superior Algorithm.

IV. DISCUSSION

From the comprehensive review of mutation strategies of BA, it can be seen that by improving and enhancing the BA following benefits can be achieved.

- a) higher convergence ratio
- b) better accuracy
- c) maximum robustness

From Table II, the results acquired from the literature show that by enhancing the BA with mutation techniques advances the features of standard BA for higher convergence ratio, better accuracy, and maximum robustness

V. CONCLUSION

BA has been widely used in various areas as an approach to solve real world nonlinear complex optimization problems. BA yet needed extreme inspection to enhance the performance of BA and researcher have suggested various BA variants for it. Table II elaborates the research participation presented.

The paper gave the detail on mutation strategies for different BA approaches utilized to resolve the local minima problem of premature convergence problem to attain the best results. We tried to give a survey on different mutation strategies and analyzed each mutation technique separately. With proper rate of growth in the research area, it is expected that more work possible in coming few years. We anticipated that this survey will provoke addition attention for these problems and major research will excite the elementary insight of how BA mutation strategies enhance the performance of standard BA. We are confident that kind of understanding may encourage the BA researchers for better awareness about a particular BA, to enhance it or to devise a new one.

REFERENCES

- [1] K. Deb, "Multi-objective optimization," in Search methodologies, Springer, Boston, MA., 2014, pp. 403-449.
- [2] Asma CHAKRI, Rabia KHELIF, Mohamed BENOURET, and Xin-She YANG, "New directional bat algorithm for continuous optimization problems," Expert Systems with Applications, vol. 69, pp. 159-175, 2017.
- [3] Ahmed Fouad Ali, "Accelerated Bat Algorithm for Solving Integer Programming Problems," Egyptian Computer Science Journal, vol. 39, pp. 25-40, 2015.
- [4] C. Koliass, G. Kambourakis, and M. Maragoudakis, "Swarm intelligence in intrusion detection: A survey," SciVersa ScienceDirect, vol. 30, no. 8, pp. 625-642, 2011.
- [5] Christian Blum and Xiaodong Li, "Swarm Intelligence in Optimization," in Swarm Intelligence., 2008, pp. 43-85.
- [6] Xin-She Yang, "Bat Algorithm for Multi-objective Optimisation," Int. J. Bio-Inspired Computation, Vol. 3, No. 5, pp.267-27, vol. 3, no. 5, pp. 267-274, 2012.
- [7] Jiann-Hong Lin, Chao-Wei Chou, Chong-Hong Yang, and Hsien-Leing Tsai, "A Chaotic Levy Flight Bat Algorithm for Parameter Estimation in Nonlinear Dynamic Biological Systems," Journal of Computer and Information Technology, vol. 2, no. 2, pp. 56-63, 2011.
- [8] Iztok Fister Jr., Dusan Fister, and Xin-She Yang, "A Hybrid Bat Algorithm," arXiv preprint arXiv:1303.6310, 2013.
- [9] Iztok Fister Jr., Simon Fong, Janez Brest, and Iztok Fister, "A Novel Hybrid Self-Adaptive Bat Algorithm," The Scientific World Journal, 2014.
- [10] Nazir Mohd Nawi, M. Z. Rehman, Abdullah Khan, Haruna Chiroma, and Tutut Herawan, "A Modified Bat Algorithm Based on Gaussian Distribution for Solving Optimization Problem," Journal of Computational and Theoretical Nanoscience, vol. 13, no. 1, pp. 706-714, 2016.
- [11] Padmavathi Kora and Sri Ramakrishna Kalva, "Improved Bat algorithm for the detection of myocardial infarction," SpringerPlus, vol. 4, no. 1, p. 666, 2015.
- [12] Xian-Bing Meng, X.Z. Gao, Yu Liu, and Hengzhen Zhang, "A novel bat algorithm with habitat selection and Doppler effect in echoes for optimization," Expert System with Applications, vol. 42, no. 1, pp. 6350-6364, 2015.
- [13] G. G. Wang, B. Chang, and Z. Zhang, "A multi-swarm bat algorithm for global optimization," in Evolutionary Computation (CEC), 2015 IEEE Congress on IEEE, 2015, pp. 480-485.
- [14] Yongquan Zhou, Jian Xie, Liangliang Li, and Mingzhi Ma, "Cloud Model Bat Algorithm," The Scientific World Journal, vol. 2014, 2014.
- [15] Gai-Ge Wang, Bao Chang, and Zhaojun Zhang, "A Multi-Swarm Bat Algorithm for Global optimization," in Evolutionary Computation (CEC), 2015 IEEE Congress on., 2015, pp. 480-485.
- [16] Xin-She Yang, "Review of Metaheuristics and Generalized Evolutionary Walk Algorithm," Int. J. Bio-Inspired Computation, vol. 3, no. 2, pp. 77-84, 2011.
- [17] Xin-She Yang and Amir Hossein Gandomi, "Bat algorithm: a novel approach for global engineering optimization," Engineering Computations, vol. 29, no. 5, pp. 464-483, 2012.
- [18] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Xin-She Yang, "Binary bat algorithm," Neural Computing and Applications, vol. 25, no. 3-4, pp. 663-681, 2014.
- [19] Suman Kumar Saha, Rajib Kar, Durbadal Mandal, Sakti Prasad Ghoshal, and Vivekananda Mukherjee, "A new design method using opposition-based BAT algorithm for IIR system identification problem," Int. J. Bio-Inspired Computation, vol. 5, no. 2, pp. 99-132, 2013.
- [20] Xin-She Yang, "Bat Algorithm: Literature Review and Applications," Int. J. Bio-Inspired Computation, vol. 5, no. 3, pp. 141-149, 2013.
- [21] Selim Yilmaz and Ecir U. Kucuksille, "Improved Bat Algorithm (IBA) on Continuous Optimization Problems," Lecture Notes on Software Engineering, vol. 1, no. 3, p. 279, 2013.
- [22] Leandro Dos Santos Coelho et al., "Bat-Inspired Optimization Approach Applied to Jiles-Atherton Hysteresis Parameters Tuning," in Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation., 2014, pp. 1455-1456.
- [23] Adis Alihodzic and Milan Tuba, "Improved Hybridized Bat Algorithm for Global Numerical Optimization," in Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on., 2014, pp. 57-62.
- [24] Xingjuan Cai, Lei Wang, Qi Kang, and Qidi Wu, "Bat algorithm with Gaussian walk," Int. J. Bio-Inspired Computation, vol. 6, no. 3, pp. 166-173, 2014.
- [25] Amir H. Gandomi and Xin-She Yang, "Chaotic bat algorithm," Journal of Computational Science, vol. 5, no. 2, pp. 224-232, 2014.
- [26] Xin-She Yang, "Swarm Intelligence Based Algorithms: A Critical Analysis," Evolutionary Intelligence, vol. 7, no. 1, pp. 17-8, 2014.
- [27] Liangliang Li and Yongquan Zhou, "A novel complex-valued bat algorithm," Neural Computing and Applications, vol. 25, no. 6, pp. 1369-1381, 2014.
- [28] Homayun Afrabandpey, Meysam Ghaffari, Abdolreza Mirzaei, and Mehran Safayani, "A Novel Bat Algorithm Based on Chaos for Optimization Tasks," in Intelligent Systems (ICIS), 2014 Iranian Conference on., 2014, pp. 1-6.
- [29] Sara Sabba and Salim Chikhi, "A discrete binary version of bat algorithm for multidimensional knapsack problem," Int. J. Bio-Inspired Computation, vol. 6, no. 2, pp. 140-152, 2014.
- [30] Md. Wasi Ul Kabir, Nazmus Sakib, Syed Mustafizur, and Mohammad Shafiul, "A Novel Adaptive Bat Algorithm to Control Explorations and Exploitations for Continuous Optimization Problems," International Journal of Computer Applications, vol. 94, no. 13, pp. 15-20, 2014.
- [31] Md. Wasi Ul Kabir and Mohammad Shafiul Alam, "Bat Algorithm with Self-adaptive Mutation: A Comparative Study on Numerical

- Optimization Problems," International Journal of Computer Applications, vol. 100, no. 10, pp. 7-13, 2014.
- [32] A. Rezaee Jordehi, "Chaotic Bat Swarm Optimisation (CBSO)," Applied Soft Computing, vol. 26, pp. 523-530, 2015.
- [33] Najmeh Sadat Jaddi, Salwani Abdullah, and Abdul Razak Hamdan, "Optimization of neural network model using modified bat-inspired algorithm," Applied Soft Computing, vol. 37, pp. 71-86, 2015.
- [34] Iztok Fister Jr. Iztok Fister and Xin-She Yang, "Towards the Development of a Parameter-free Bat Algorithm," in *StuCoSReC: Proceedings of the 2015 2nd Student Computer Science Research Conference.*, 2015, pp. 31-34.
- [35] Laamari Mohamed Amine and Kamel Nadjet, "A Multi-objective Binary Bat Algorithm," in *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication.*, 2015, p. 75.
- [36] Zhihua Cui, Feixiang Li, and Qi Kang, "Bat Algorithm with Inertia Weight," in *Chinese Automation Congress (CAC)*, 2015., 2015, pp. 792-796.
- [37] Yimin Deng and Haibin Duan, "Chaotic Mutated Bat Algorithm Optimized Edge Potential Function For Target Matching," in *Industrial Electronics and Applications (ICIEA)*, 2015 IEEE 10th Conference on., 2015, pp. 1049-1053.
- [38] Jonathan Pérez, Fevrier Valdez, and Oscar Castillo, "Proposed Augmentation of the Bat Algorithm using fuzzy logic for dynamic parameter adaptation," in *Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)*, 2015 Annual Conference of the North American., 2015, pp. 1-6.
- [39] Selim Yilmaz and Ecir U. Küçükşille, "A new modification approach on bat algorithm for solving optimization problems," Applied Soft Computing, vol. 28, pp. 259-275, 2015.
- [40] Xianbing Meng, X. Z. Gao, and Yu Liu, "A Novel Hybrid Bat Algorithm with Differential Evolution Strategy for Constrained Optimization," *International Journal of Hybrid Information Technology*, vol. 8, no. 1, pp. 383-396, 2015.
- [41] Jinfeng Wang, Xiaoliang Fan, Ailin Zhao, and Mingqiang Yang, "A Hybrid Bat Algorithm for Process Planning Problem," *ScienceDirect*, vol. 48, no. 3, pp. 1708-1713, 2015.
- [42] Apurv Shukla, "A modified Bat Algorithm for the Quadratic Assignment Problem," in *Evolutionary Computation (CEC)*, 2015 IEEE Congress on., 2015, pp. 486-490.
- [43] Gai-Ge Wang, Mei Lu, and Xiang-Jun Zhao, "An Improved Bat Algorithm with Variable Neighborhood Search for Global Optimization," in *Evolutionary Computation (CEC)*, 2016 IEEE Congress on., 2016, pp. 1773-1778.
- [44] Jonathan Perez, Fevrier Valdez, Oscar Castillo, and Olympia Roeva, "Bat Algorithm with Parameter Adaptation using Interval Type-2 Fuzzy Logic for Benchmark Mathematical Functions," in *Intelligent Systems (IS)*, 2016 IEEE 8th International Conference on., 2016, pp. 120-127.
- [45] Hema Banati and Reshu Chaudhary, "Enhanced Shuffled Bat Algorithm (EShBAT)," in *Advances in Computing, Communications and Informatics (ICACCI)*, 2016 International Conference on., 2016, pp. 731-738.
- [46] Sumitha Manoj, S. Ranjitha, and Dr. Suresh H N, "Hybrid BAT-PSO Optimization Techniques for Image Registration," in *Electrical, Electronics, and Optimization Techniques (ICEEOT)*, International Conference on., 2016, pp. 3590-3596.
- [47] Liu Yi, Diao Xingchun, Cao Jianjun, and Zhang Bin, "Local Enhanced Catfish Bat Algorithm," in *Robots & Intelligent System (ICRIS)*, 2016 International Conference on., 2016, pp. 238-245.
- [48] Aadil Latif, Ishtiaq Ahmad, Peter Palensky, and Wolfgang Gawlik, "Multi-Objective Reactive Power Dispatch in Distribution Networks using Modified Bat Algorithm," in *Proceedings of the IEEE Green Energy and System Conference (IGESC)*, 2016.
- [49] Sinha Sheikh Abdullah, Shabnam Sharma, Kiran Jyoti, and U. S. Pandey, "Review of Recent load Balancing Techniques in Cloud Computing and BAT Variants," in *Proceedings of the International Conference on Computing for Sustainable Global Development(INDIACom)*, 2016.
- [50] Messaoudi Imane and Kamel Nadje, "Hybrid Bat algorithm for overlapping community detection," *ScienceDirect*, vol. 49, no. 12, pp. 1454-1459, 2016.
- [51] Sonia Goyal and Manjeet Singh Patterh, "Modified Bat Algorithm for Localization of Wireless Sensor Network," *Wireless Pers Commun*, vol. 86, no. 2, pp. 657-670, 2016.
- [52] Jigyasha Sharma and Annappa B, "Community Detection Using Meta-heuristic Approach: Bat Algorithm Variants," in *Contemporary Computing (IC3)*, 2016 Ninth International Conference on., 2016, pp. 1-7.
- [53] N. Mohan, R. Sivaraj, and R. Devi Priya, "A Comprehensive Review of BAT Algorithm and its Applications to various Optimization Problems," *Asian Journal of Research in Social Sciences and Humanities*, vol. 6, no. 11, pp. 676-690, 2016.
- [54] Rohit Salgotra and Urvinder Singh, "A novel bat flower pollination algorithm for synthesis of linear antenna arrays," *Neural Comput & Applic*, 2016.
- [55] Xingjuan Cai, Xiao-zhi Gao, and Yu Xue, "Improved bat algorithm with optimal forage strategy and random disturbance strategy," *International Journal of Bio-Inspired Computation*, vol. 8, no. 4, pp. 205-214, 2016.
- [56] Deepika Singh, Rohit Salgotra, and Urvinder Singh, "A Novel Modified Bat Algorithm for Global Optimization," in *Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017 International Conference on., 2017, pp. 1-5.
- [57] Jinle Li1, Huazhong Wang1, and Bingyong Yan1, "Application of Velocity Adaptive Shuffled Frog Leaping Bat Algorithm in ICS Intrusion Detection," in *Control And Decision Conference (CCDC)*, 2017 29th Chinese., 2017, pp. 3630-3635.
- [58] Siqing Sheng and Jingjing Zhang, "Capacity configuration optimisation for stand-alone micro-grid based on an improved binary bat algorithm," *The Journal of Engineering*, vol. 2017, no. 13, pp. 22083-2087, 2017.
- [59] Mohamed A. Tawhid and Ahmed F. Ali, "Multi-directional bat algorithm for solving unconstrained optimization problems," *OPSEARCH*, vol. 54, no. 4, pp. 684-705, 2017.
- [60] Asma Chakri, Xin-She Yang, Rabia Khelif, and Mohamed Benouaret, "Reliability based-design optimization using the directional bat algorithm," *Neural Computing and Applications*, pp. 1-22, 2018.
- [61] W. H. Bangyal, J. Ahmad, H. T. Rauf, and S. Pervaiz, "An Improved Bat Algorithm based on Novel Initialization Technique for Global Optimization Problem," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 9, no. 7, pp. 158-166, August 2018.

TABLE II. YEAR WISE RELATED WORK WITH DETAILED OBSERVATION

Review						
S.No	Authors	Year	Paper Title	Modification Techniques	OBSERVATION	Results Validation
1	Xin-She Yang	2011	Review of Metaheuristics and Generalized Evolutionary Walk Algorithm	Generalized Evolutionary walk algorithm (GEWA)	Initial version of BA	-
2	Jiann-Horng Li	2011	A Chaotic Levy Flight Bat Algorithm for Parameter Estimation in Nonlinear Dynamic Biological Systems	Chaotic Levy Fight Bat Algorithm	Standard Bat Algorithm	Effective
3	Jiawei Zhang	2012	Image Matching using a Bat Algorithm with Mutation	Bat Algorithm with Mutation (BAM)	Simple BAT and other optimization algorithms	superior
4	Xin-She Yang	2012	Bat algorithm: a novel approach for global engineering optimization	new BA	Seven benchmark engineering problems	significant results
5	Xin-She Yang	2012	Bat Algorithm for Multi-objective Optimisation	Multiobjective Bat Algorithm (MOBA)	sub-set of test functions and multiobjective design problems	Efficient
6	Galge Wang	2013	A Novel Hybrid Bat Algorithm with Harmony Search for Global Numerical Optimization	novel robust hybrid metaheuristic optimization method (HS/BA)	Fourteen test functions	Outperforms
7	Iztok Fister Jr	2013	A Hybrid Bat Algorithm	Hybrid Bat Algorithm (HBA)	Standard Benchmark Test Functions	Improved
8	Seyedali Mirjalili	2013	Binary Bat Algorithm	Binary Bat Algorithm (BBA)	Twenty two standard functions, Binary GA and PSO	Outstanding Results
9	Suman Kumar Saha	2013	A new design method using opposition-based BAT algorithm for IIR system identification problem	Opposition-Based Bat Algorithm (OBA)	Standard Bat Algorithm	Effective
10	Xin-She Yang	2013	Bat Algorithm: Literature Review and Applications	Review	Standard Algorithms	significant results
11	Selim Yilmaz	2013	Improved Bat Algorithm (IBA) on Continuous Optimization Problems	Three modifications in standard BAT	Ten standard functions	Better in quality
12	S.Yilmaz	2014	Modified Bat Algorithm	Modified Bat Algorithm (MBA)	fifteen standard Test functions	Quality solution
13	Leandro dos Santos Coelho	2014	Bat-Inspired Optimization Approach Applied to Jiles-Atherton Hysteresis Parameters Tuning	Hybrid BA with a mutation style operator from differential evolution and beta probability distribution (BADEBD)	Standard Bat Algorithm	Better in performance
14	Adis Alihodzic	2014	Improved Hybridized Bat Algorithm for Global Numerical Optimization	Modified Bat-inspired differential evolution	Five standard functions	Outperforms
15	Iztok Fister Jr	2014	A Novel Hybrid Self-Adaptive Bat Algorithm	Hybrid Self-Adaptive Bat Algorithm (HSABA)	Standard Algorithms	HSABA works adequately
16	Osama Abdel-Raouf	2014	An Improved Chaotic Bat Algorithm for Solving Integer Programming Problem	Improved Chaotic Bat Algorithm (IBACH)	Traditional BA, PSO and Other Search Algorithms	Works with less computational time
17	XingjuanCai	2014	Bat algorithm with Gaussian walk	BAT Algorithm with Gaussian Walk (BAGW)	Four benchmark functions	Excellent
18	Amir H. Gandomi	2014	Chaotic Bat Algorithm	BAT algorithm with chaos	Standard functions	Outperforms
19	Xin-She Yang	2014	Swarm Intelligence Based Algorithms: A Critical Analysis	Critical analysis on SI	Self-organization, Markov chain framework and dynamic system	
20	Liangliang Li	2014	A novel complex-valued bat algorithm	Complex-Valued Bat Algorithm	PSO and fourteen test functions	Feasible results
21	Homayun Afrabandpey	2014	A Novel Bat Algorithm Based on Chaos for Optimization Tasks	Chaotic Bat Algorithm (CBA)	Standard Test functions and traditional BAT	Efficient
22	Sara Sabba	2014	A discrete binary version of bat algorithm for multidimensional knapsack problem	Discrete Binary Bat Algorithm (BinBA)	Multidimensional Knapsack problem	Promising results
23	Md. Wasi Ul Kabir	2014	A Novel Adaptive Bat Algorithm to Control Explorations and Exploitations for Continuous Optimization Problems	Novel Adaptive Bat Algorithm (NABA)	Test functions nad Traditional BAT	Compatible
24	Md. Wasi Ul	2014	Bat Algorithm with Self-adaptive	Self-Adaptive Bat	Uni-model and multi-	Outperforms

	Kabir		Mutation: A Comparative Study on Numerical Optimization Problems	Algorithm (BA-SAM)	model optimization test problems	
25	Yongquan Zhou	2014	Cloud Model Bat Algorithm	Cloud Model Bat Algorithm (CBA)	Function optimization	Good performance
26	A. Rezaee Jordahi	2015	Chaotic Bat Swarm Optimisation (CBSO)	Chaotic Bat Swarm Optimisation (CBSO)	Various test functions	Good in quality
27	Najmeh Sadat Jaddi	2015	Optimization of neural network model using modified bat-inspired algorithm	Modified Bat Algorithm	Six classifications and two standard time series datasets	Satisfactory results
28	Iztok Fister Jr	2015	Towards the Development of a Parameter-free Bat Algorithm	Advanced BAT	Standard functions	Showed good capability
29	Laamari Mohamed Amine	2015	A Multi-objective Binary Bat Algorithm	Multi-Objective Binary Bat Algorithm (MBBA)	non-dominated sorting genetic algorithm 2 (NSGA-2)	Better in performance
30	Gai-Ge Wang	2015	A Multi-Swarm Bat Algorithm for Global optimization	Multi-Swarm Bat Algorithm (MBA)	Satndard BAT and sixteen standard functions	Significant in performance
31	Zihuha Cui	2015	Bat Algorithm with Inertia Weight	New parameter with inertia weight	CEC2013 standard problems	Validate
32	Yimin Deng	2015	Chaotic Mutated Bat Algorithm Optimized Edge Potential Function For Target Matching	Chaotic Mutated Bat Algorithm (CMBA) Optimized Edge Potential Function (EPF)	Real world applications	works efficiently
33	Jonathan Pérez	2015	Proposed Augmentation of the Bat Algorithm using fuzzy logic for dynamic parameter adaptation	Bat Algorithm improvement by using fuzzy logic for dynamic parameter	Satndard BAT and GA	Outstanding Results
34		2015	A new modification approach on Bat Algorithm for solving optimization problem	Enhanced Bat Algorithm (EBA)	Standard Functions	Excellent Results
35	Xian-Bing Meng	2015	A novel bat algorithm with habitat selection and Doppler effect in echoes for optimization	Noval Bat Algorithm (NBA)	Twenty Standard functions and four real life engineering applications	Superior
36	Xianbing Meng	2015	A Novel Hybrid Bat Algorithm with Differential Evolution Strategy for Constrained Optimization	Hybrid Bat Algorithm with Differential Evolution Strategy (BADE)	Nine standard functions and three real life engineering design problems	Effective Results
37	Ahmed Fouad Ali	2015	Accelerated Bat Algorithm for Solving Integer Programming Problem	Accelrated Bat Algorithm (ABATA)	four standard algorithms and seven integer programming problems	works in less computational time
38	Jinfeng Wang	2015	A Hybrid Bat Algorithm for Process Planning Problem	Hybrid BAT	Two standard algorithm	Satisfactory Solution
39	Padmavathi Kora	2015	Improved Bat algorithm for the detection of myocardial infarction	Improved Bat	Four benchmark functions	Better in performance
40	Apurv Shukla	2015	A modified Bat Algorithm for the Quadratic Assignment Problem	Modified Bat	Traditional BA, PSO and Other Search Algorithms	works efficiently
41	Gai-Ge Wang	2016	An Improved Bat Algorithm with Variable Neighborhood Search for Global Optimization	Variable Neighbourhood Bat Algorithm (VNBA)	Sixteen Standard Test Functions	Outperforms
42	Jonathan Perez	2016	Bat Algorithm with Parameter Adaptation using Interval Type-2 Fuzzy Logic for Benchmark Mathematical Functions	a new methodology of Dynamic behavior of BAT parameters	Type-1 fuzzy Logic	Best results
43	Hema Banati	2016	Enhanced Shuffled Bat Algorithm (EShBAT)	Enhanced Shuffled Bat Algorithm (EShBAT)	Thirty Test Functions	Superior results
44	Sumitha Manoj	2016	Hybrid BAT-PSO Optimization Techniques for Image Registration	BAT-PSO	Ten standard functions	Better in performance
45	Liu Yi	2016	Local Enhanced Catfish Bat Algorithm	Local Enhanced Catfish Bat Algorithm (LECBA)	Modified BAT	Excellent Results
46	Aadil Latif	2016	Multi-Objective Reactive Power Dispatch in Distribution Networks using Modified Bat Algorithm	Modified BA	Standard bench mark functions	Outperforms
47	Shabnam Sharma	2016	Research & Analysis of Advancement in BAT Algorithm	Reviewof relate work	Un-constrained problems	Best results
48	Sinha Sheikh Abdullah	2016	Review of Recent load Balancing Techniques in Cloud Computing and BAT Variants	V-BAT	Standard functions	works efficiently
49	Messaoudi	2016	Hybrid Bat algorithm for overlapping	Hybrid Bat Algorithm	Two standard	Superior

	Imane		community detection		algorithm	
50	Sonia Goyal	2016	Modified Bat Algorithm for Localization of Wireless Sensor Network	Modified BA	Ten standard functions	Better
51	Jigyasha Sharma	2016	Community Detection Using Meta-heuristic Approach: Bat Algorithm Variants	Four new variants of BAT	fast greedy modularity optimization, Grivan and Newman, Reichardt and Bornholdt, Ronhovde and Nussinoy and spectral clustering	Promising results
52	N. Mohan	2016	A Comprehensive Review of BAT Algorithm and its Applications to various Optimization Problems	Survey on BAT applications	Constrained problems	works efficiently
53	Nazir Mohd Nawi	2016	A Modified Bat Algorithm Based on Gaussian Distribution for Solving Optimization Problem	Improved Bat Algorithm with Guassian distribution random walk (BAGD)	Ten standard test functions along with six meta heuristic algorithms	Better in performance
54	Rohit Salgotra	2016	A novel bat flower pollination algorithm for synthesis of linear antenna arrays	Bat Flower Pollination (BFP)	Ten test functions and other famous techniques	Superior results
55	Xingjuan Cai	2016	Improved bat algorithm with optimal forage strategy and random disturbance strategy	Optimal Forage Strategy	CEC2013 test suite along with four evolutionary algorithms	Effective
56	Deepika Singh	2017	A Novel Modified Bat Algorithm for Global Optimization	Modified Bat Algorithm (MBA)	Standard benchmark problems and state-of-art algorithms	Better
57	Jinle Li	2017	Application of Velocity Adaptive Shuffled Frog Leaping Bat Algorithm in ICS Intrusion Detection	Velocity Adaptive Shuffled Frog Leaping BAT Algorithm (VASFLBA)	Standard Functions	Better performance
58	Siquing Sheng	2017	Capacity configuration optimisation for stand-alone micro-grid based on an improved binary bat algorithm	Improved Binary Bat Algorithm (IBBA)	PSO and GA	Better in stability
59	Fábio A. P. Paiva	2017	Modified Bat Algorithm With Cauchy Mutation and Elite Opposition-Based Learning	Modified Bat having Cauchy mutation and elite opposition Based learning	Four benchmark functions	Excellent
60	Mohamed A. Tawhid	2017	Multi-directional bat algorithm for solving unconstrained optimization problems	Multi-Directional Bat Algorithm (MDBAT)	Eight standard algorithm and sixteen global optimization problems	Outperforms
61	Asma CHAKRI	2017	New directional bat algorithm for continuous optimization problems	Directional BAT Algorithm (dBA)	BAT variations, ten standard algorithm and functions from CEC'2005 standard suite	Better
62	Asma CHAKRI	2018	Reliability based-design optimization using the directional bat algorithm	Novel resolution method belongs to Directional Bat algorithm	Various engineering problems	Solve RDBO problems
63	Waqas Haider Bangyal	2018	An Improved Bat Algorithm based on Novel Initialization Technique for Global Optimization Problem	Improved Standard BA (I-BAT)	Standard Functions	Superior results

Arabic Chatbots: A Survey

Sarah AlHumoud, Asma Al Wazrah, Wafa Aldamegh
College of Computer and Information Science
Al-Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

Abstract—A Chatbot is a programmed entity that handles human-like conversations between an artificial agent and humans. This conversation has attracted the attention of researchers who are interested in the interaction between humans and machines to make the conversation more rational and hence pass the Turing test. The available research done in the field of Arabic chatbots is comparably scarce. This paper presents a review of the published Arabic chatbots studies to identify the gap of knowledge and to highlight the areas that needs more study and research. This study concluded the rarity of available research on Arabic chatbots and that all available works are retrieval based.

Keywords—Artificial intelligence; Arabic chatbot; conversational agent; ArabChat; human-machine interaction; utterance

I. INTRODUCTION

Artificial Intelligence (AI) is focused on the learning processes. The idea of using a human language to communicate with computers is holding merit to AI. A chatbot or a chat-agent is an intelligent conversation agent which interacts with human users via natural language and emulates human conversation. This area has attracted more interest from both research and industry fields in the past few years [1]. The first chatbot was developed in Massachusetts Institute of Technology (MIT), where Weizenbaum implemented the ELIZA chatbot to emulate a psycho-therapist in 1966 [2].

Nowadays, a variety of chatbots are available online to serve in different domains ranging from customer service and information acquisition to entertainment where primarily users react with these applications to make a small conversation. They extend from unpretentious systems which extract answers from datasets when they match specific keywords to more advanced ones which utilize Natural Language Processing (NLP) techniques. A chatbot could be programmed to serve almost any human language. Although the research on English chatbots is diffuses heavily, there is a Scarcity in the Arabic chatbots due to difficulties in the Arabic language. In this paper, we present a survey on Arabic chatbots. The number of online Arabic users have increased which motivated to build Arabic chatbots.

Processing Arabic language texts have a lot of challenges [3] such as rich morphology, the high degree of ambiguity, orthographic variations, and the existence of multiple dialects. Moreover, written Arabic text can be classified into three categories. First, Classical Arabic (CAL) or Quranic Arabic in the Holy Quran. Second, the Modern Standard Arabic (MSA) that is the official language in the Arab world and used in a

formal written and spoken forms in mediums such as news, education, and literature [4]. The third category is Dialectal Arabic (DA), which is used daily in spoken and written personal communication and in informal settings, where each country and region has its own dialect [2].

The rest of the paper is organized as follows: Section II III presents the background. Section III presents the survey methodology. Section IV discusses the Arabic chatbots researches. Section V presents the conclusion.

II. BACKGROUND

There are generally three components of Chabot [5]: the interface, which interacts with user's input and output, the Knowledge Base or brain, which include the content of the conversation and keep truck of the domain, and the Conversation Engine, which manages the semantic context of the conversation.

There are two types of dataset models, which represent the knowledge source type in chatbots: the retrieval-based model and the generative-based model. In the retrieval-based model, a chatbot uses a pool of predefined responses and employs a type of heuristics to select the proper response to the input, but it may not be applicable when there is no existing predefined response. In the generative-based model, a chatbot uses a set of the techniques for generating new responses and could utilize predefined responses as well using deep learning and neural network (NN). In the following, we will discuss the literature based on the different techniques used, the length of the conversation, the domain of conversation, and the dataset model.

In the retrieval-based model, there are common techniques to build the conversational agent, using pattern matching, Artificial Intelligence Markup Language (AIML), Ontologies, Parsing, Markov Chain Model, and ChatScript. While in the generative-based model, the different techniques are neural network and deep learning techniques. Seq2seq based on neural network will be introduced.

Pattern matching is used mainly in the question/answer chatbots, where the system matches the input with a predefined structure to create a response. AIML is widely used in chatbot design, it is a language derived of XML. It represents the knowledge as objects, consisting of topics and categories. The AIML pattern consists of words having letters and numerals but no special characters or spaces [6]. Many chatbot applications have emerged in English. A.L.I.C.E. [7] is a retrieval-based model and it is using advanced pattern matching and AIML approach. It is the first AIML-based

chatbot and won the Loebner Prize in 2000, 2001 and 2004. A.L.I.C.E is a supervised learning and it is based on categories containing a pattern, and a template for the response. Category patterns are matched to find the most appropriate response to a user input. AIML tags provided for consideration of context, conditional branching to produce new responses. Some Arabic studies are applied to A.L.I.C.E.

Ontology [8] is a set of interconnected hierarchy classes. The knowledge base can be described as a graph that contains classes, each class describes the concepts and the properties. The classes that have a logical relationship are also connected, and use these relationships to imply new statements (reasoning). Examples of ontologies are OpenCyc and Wordnet.

Textual Parsing [9] is a method which converts the text into a set of words (lexical parsing) to determine its grammatical structure. After the tree is built from these words, the lexical structure can be then checked if it forms the rule of the language (syntactical parsing). The latter parsers are getting more complex using natural language processing.

The Markov Chain Model [8] depend on the probability of occurrence of a word or letter in the input text, this method helps in building responses that are probabilistically more suitable and hence more correct. For example, if an input text is "xyyyzxyzyzyzy", then the Markov model of order 0 predicts that letter "x" occurs with a probability $2/13$. The Markov model of order 1 predicts the fixed probability for every letter depends on the previous letter.

ChatScript [8], [9] aims to be easier to maintain than AIML by focusing on better syntax, it fixes the zero-word matching problems. The Chatscript first finds the best topic that matches the user query string and executes a rule in that topic. Rather than using separated categories for each word as in AIML, Chatscript uses 'concepts' to merge similar words with meanings or parts of speech. Suzette (written in ChatScript) won the 2010 Loebner Prize.

On the other hand, there is a common technique to build a generative-based model chatbot using a Recurrent Neural Network (RNN). Seq2seq model [10] is an encoder-decoder model that uses RNN and it is primarily used for translating from one language to some other language, but in the context of chatbots, the input is translated to a response. The seq2seq model is composed of two main RNNs, an encoder RNN which takes the input sequence and encapsulates the information into a fixed representation one cell at time, and a decoder RNN which take that representation, and generates a variable length text that best responds to it also one cell at time. Seq2seq encodes only the important information in the sequence and convert a sequence of symbols into a fixed size feature vector. The cell used in RNN is long short-term memory (LSTM), It allows the cells to remember what information needs to be remembered or updated from the previous cells [10]. On scanning the literature, no studies were found to apply a neural network in Arabic chatbot design.

In addition to the technique used and the type of data being processed, the length of the conversation and the domain of conversation, as well as the dataset model are considered

aspects in this survey and will be discussed. The conversation length is classified into short and long conversation. A short conversation is a single response produced for a single input such as a question/ answer conversation. Where a long conversation indicates that a large amount of information is exchanged during the conversation lifetime, and this information is tracked and may be present in the output. On the conversation domain, chatbots are classified into two types, closed and open. The closed domain is designed to serve a specific purpose, where the knowledge that is required to generate a suitable response to an input is limited. While the open domain is like human's conversation, the domain may change with the time, supporting more than one conversation domain.

III. METHODOLOGY

The research survey methodology consists of scanning different literature databases. A similar methodology is used in the literature review here [11]. The literature collection was done in highly cited computer science libraries like: IEEE, ACM, Springer, Science Direct and Google Scholar. The search was done using ten keywords coupled with the keyword 'Arabic'. Those keywords are 'chatbot', 'chatterbot', 'ArabChat', 'chat agent', 'interactive agent', 'conversational agent', 'conversational robot', 'artificial conversational', 'dialogue', and 'utterance'. The result consisted of 184 papers as shown in Fig. 1. Those papers collected from 2004 to 2017 and evaluated according to the title and the abstract of the paper, eliminating the papers that do not present an implementation of an Arabic chatbot. After evaluation and elimination, we found that there are fourteen papers present Arabic chatbot application, which was from the IEEE and Springer libraries and Google Scholar.

IV. ARABIC CHATBOT RELATED WORK

Although there are available developed Arabic chatbots applications, the research available on Arabic chatbots is limited. Some examples on the former are services as Al-Haj Bot, Rammas, Msa3ed, Theyabi, El-Kahwagy, and others provide an Arabic chatbot application developed for a commercial purpose. Also, there are platforms that provide developers with coding facility and aid such as Watson by IBM, Messenger Bot by Facebook, Telegram Bot, PandoraBot. These platforms and applications are excluded from this review for the lack of published research on them, making it difficult to analyze and compare fairly. The purpose of this survey is to highlight state of the art Arabic chatbot research.

The fourteen collected papers present twelve different Arabic chatbot applications, that classified and evaluated in a manner similar to what is applied here [12], [8], [9]. Based on the data type a chatbot processes, it is classified into two categories, text and speech conversation chatbot. The classification relies on the chatbot input and output interaction. In each category, chatbots are classified based on the implementation technique into two subcategories namely, pattern matching and AIML approach. In addition to that, we will discuss chatbot aspects such as the length of the conversation in terms of interaction duration, the domain of conversation as topics domain that chatbot can interact with, and the dataset model of the chatbot.

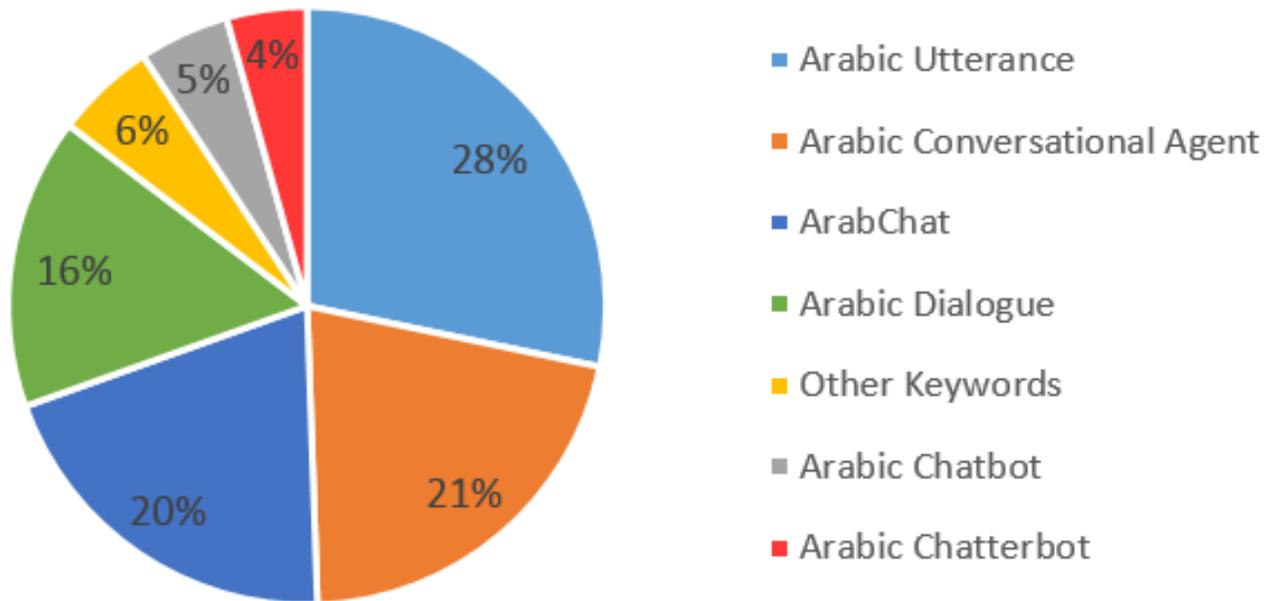


Fig. 1. Total Results from Literature Databases For Search Keywords.

A. Text Conversational Chatbot

In this category, the interaction with the chatbot conversation is through textual input and output. The chatbot in this category is classified based on the implementation technique into two subcategories, those are AIML and pattern match approach.

1) *AIML approach*: Among the earliest research studies on Arabic chatbot application in the collected related work is Quran chatbot [13] by Shawar and Atwell. It is a chatbot on the Quran Islamic holy book. The Quran contains 6236 verses or 'Ayahs' and 114 'Surahs' which is a set of verses. The format of the user inputs are Arabic words with 'Tashkil' or diacritics, that is used as phonetic guides. The chatbot replies by finding the 'Ayahs' from the Quran that contain the user's input. The nature of Quran text is non-conversational, a Java program is developed to adopt a learning process. The learning process was based on the most significant word of the 'Ayha' that represents the category in the AIML file and template is the Ayah. Arabic AIML file is generated by the Java program. The conversation length is short since the chatbot responses with a single response to a single user input. The domain of the conversation is limited by the content of the Quran. From that, the sources of the chatbot dataset are retrieval-based. The interaction and response of the chatbot are limited by the pool of the most significant word of the 'Ayhas' that are extracted by the java program.

The Quran14-114 chatbot by Shawar and Atwell [6] is a version of the ALICE chatbot [14] added to the Java program to interact as the Quran chatbot. The conversation in the chatbot is short. The user inputs a question or a statement in English, and the chatbot responses with one or more appropriate 'Surahs' and 'Ayahs' from the Quran in both

English and Arabic. The Java program reads the Quran text from a corpus and converts it to the AIML format to be used by ALICE chatbot. The domain of the conversation is closed based on the content of the Quran in both the English and Arabic languages. Also, this is a retrieval-based model as the dataset is limited by the content of the Quran. The challenge in this work is to show how ALICE chatbot adapted to learn from non-conversational text.

The Arabic Web Question Answering (QA) chatbot [2], [15] is a web interface chatbot based on an Arabic QA corpus, that was built from five different web pages with 412 Arabic question and answer. Those web pages' cover topics such as motherhood and pregnancy, dental care, fasting and health, blood disease such as cholesterol and diabetes, and blood charity. The chatbot supports more than one closed domain, and thus regarded as a closed conversation domain. The chatbot conversation is short, where the user inputs a textual question in MSA about one of the supported domains and, the chatbot responses with the answer without using sophisticated NLP. Also, a Java program was developed to convert the text corpus to create two AIML files atomic and default. The atomic file contains the questions and answers that appear in the corpus. The default file is used to guarantee that the user question is mapped to the appropriate question stored in the knowledge base. Moreover, the file is built using the first word and the most significant word approach. The first word acts as a classifier to the question and the most significant word is the least frequent in the question.

The latter is done by building questions' frequency list after applying a tokenization process to the question. The generated list contains the question's words along with their frequencies. Then, the approach extracts the two most significant words in the list, those are the two least frequent words, used as keywords to map the question to an answer. The purpose of

using the most significant word approach was to increase the rate of the expected output. The chatbot was tested by entering fifteen questions and the result was 93% correct answers. The main drawback in this model appears when the structure of the question is changed or altered from the stored in the knowledge base, then the chatbot responses with wrong answers. That happens because the chatbot does not use a heuristic to select a proper response and it is based on a direct retrieval model. Also, a success rate of to be 93% is not justifiable having a dataset of fifteen questions only.

BOTTA chatbot by Ali and Habash [3] is a female chatbot, that simulates friendly conversations with users. The chatbot supports Egyptian Arabic dialect for both input and output. BOTTA is available to the public. It simulates the English chatbot Rosie [16]. The knowledge base is made up of Rosie's AIML files set. Some of Rosie's AIML files are translated directly to Arabic, and the others are modified according to the use of the Arabic dialects. Also, for each conversation, BOTTA chatbot temporarily stores the basic information about the user such as age, gender, and nationality by asking questions yielding a conversation that is open since the chatbot can response to different topics domain. The length of the conversation is long where chatbot can response to the user based on previous information in conversation. However, it does not update the knowledge base and add new responses, so it is based on retrieval-based model. It depends on a pool of predefined responses using heuristics to response with an appropriate output. Also, the chatbot does not perform the text normalization on the user input to get the suitable response. It performs orthographic transformations, that includes correcting common spelling mistakes of the user input. With this method, BOTTA was able to resolve 85.1% of the common spelling mistakes in Arabic typing.

Table I shows a summary on the discussed textual conversation chatbots that uses AIML approach.

TABLE I. A SUMMARY OF TEXTUAL ARABIC CHATBOTS USING AIML APPROACH

Chatbot	Language	Interaction Type		Conversation Length	Domain	Dataset Model
		Input	Output			
Quran [13]	Classical Arabic	Text	Text	Short	Closed (Quran book)	Retrieval-Based
Quran14-114 [6]	English/classical Arabic	Text	Text	Short	Closed (Quran book)	Retrieval-Based
Web QA [2], [15]	Arabic (MSA)	Text	Text	Short	Closed (Medical care)	Retrieval-Based
Botta [3]	Egyptian Arabic dialect	Text	Text	Long	Open	Retrieval-Based

2) *Pattern matching approach*: Mohammad Hijjawi, Zuhair Bandar, Keeley Crockett and David Mclean [17] implemented the ArabChat, which is a conversational agent web interface. The chatbot conversation domain is closed, designed to serve the students of the Applied Science University in Jordan. The interaction between the user and the chatbot is through textual Arabic MSA language. The conversation remains ongoing until one of the conversation's parties terminates it. The ArabChat reuses the previously exchanged information during the conversation as a response to the user input, creating long conversations. The core components of the ArabChat chatbot are the scripting engine and a scripting language. The scripting engine is divided into subcomponents, that allows handling topics of the conversations. ArabChat knowledge base contains 1218 utterances, that are classified into contexts, each context contains rules. The rules consist of patterns and associated textual responses. ArabChat was tested over 174 users, the average input for each user was 7 inputs per user. The result shows that 73.56% of the inputs matched the expected output.

Enhanced ArabChat [18] is an updated version of ArabChat [17] by Hijjawi, Bandar and Crockett. This version uses extra features including Utterance Classification and Hybrid Rule. These improvements were at the engine level while some additional improvements need to be added to the scripting language and knowledge base to meet the changes needs. Utterance classification feature aims to distinguish between a question and non-question utterances. It works by adding extra keywords to the pattern of the question-based rule, to deal with keyword matching. Hybrid Rule is the second feature and it focuses on how to reply and deal with an utterance that request many topics. Although ArabChat gave a better result of Ratio of Matched Utterances to the Total (RMUT) than enhanced one due to unserious users, the manual checking gives more accurate results and showed improvement in performance. By analyzing logs manually, Enhanced ArabChat deals successfully with 82% of utterances with two topics and this ratio is decreasing when the number of topics is increased in the utterance. Using manual checking, classifying utterance shows a high percentage of question-based utterances due to three factors: the selected domain, the users' needs, that implies that they are more likely to ask rather than discuss, and difficulties to script a large number of rules.

ArabChat with classification methodology [19] is another ArabChat [17] update by Hijjawi, Bandar and Crockett. Using a new classification methodology for Arabic utterances. This new approach classifies the sentences into questions and non-questions including assertions and instructions. The benefit of applying this approach is that the number of patterns required per rule will decrease and hence increase the performance by firing the suitable rule, depending on the utterance type being a question or non-question. Different topics and list of function words have been used from domains such as politics, religion, sports, education, business and adding some synthetic non-question sentences and indirect questions. This classification is done by pre-processing the Arabic sentence into equivalent numeric tokens and then importing the tokens into a machine

learning toolkit in WEKA. In WEKA, a Decision Tree, which achieve the highest accurate classifier to be applied on the tokenized numeric dataset, is generated and then is converted into a standard IF-THEN classification rule to classify utterances.

Mobile ArabChat [20] is based on the original ArabChat [17] and it is a mobile-based conversational agent and it is also used to work as an advisor for students in Applied Science University in Amman. It is a light version of ArabChat implemented in Android. Although there are some challenges facing users in the Arab Countries such as slow and unstable internet connection and limited bandwidth, this application works even with limited Internet bandwidth. Mobile ArabChat implemented pattern matching approach based on the text. This framework consists of the same component as in ArabChat: scripting engine, scripting language and a knowledge base. Based on a subjective approach, 96% of users agree that using Mobile ArabChat via mobile is better than using the same system via desktop. However, Mobile ArabChat needs an internet connection to work.

Abdullah [4] is an Arabic Conversational Intelligent Tutoring System (CITS) that teaches children aged 10 to 12 years old essential topics about Islam. This online system can engage with students using MSA. That asking a series questions to the students, and discuss with them their answers, using Classical Arabic to give evidences from the Quran and Hadith, which is the sayings and traditions of the Prophet of Islam Muhammed. The system is using images and sound effects to interact with students and can determine the student's knowledge level and hence direct the conversation. Abdullah CITS can distinguish between the user's questions and answers. The framework is based on a Pattern Matching approach, it consists of knowledge base having subject topics, the Conversational Agent scripting language to deliver the tutorial conversation to the learners, and The Tutorial Knowledge Base to determine the level of individual student knowledge and the subject.

LANA [21] is another CITS and it was developed for children with Autism Spectrum Disorder (ASD) that are 10 to 16 years old who have reached a basic competency with the mechanics of Arabic writing to teach them topics on science using MSA. Children with ASD have difficulties in traditional learning because the teacher can't meet the need of every individual student. LANA engages children with a science tutorial delivered in MSA. It is similar to Abdullah CITS, but it offers different learning style models such as visual, auditory and kinesthetic, enabling children to practice learning skills independently based on their needs using pattern matching and short text similarity algorithm. This system also interacts with children using materials such as picture, audio, or instructions according to the user's learning style. Table II shows a brief review of the discussed related pattern matching text conversational chatbots.

TABLE II. A SUMMARY OF TEXTUAL ARABIC CHATBOTS USING PATTERN MATCHING APPROACH

Chatbot	Language	Interaction Type		Conversation Length	Domain	Dataset Model
		Input	Output			
ArabChat [17]	Arabic (MSA)	Text	Text	Long	Closed (for students of Applied Science University)	Retrieval-Based
Enhanced ArabChat [18]	Arabic MSA	Text	Text	Long	Close (for students of Applied Science University)	Retrieval-Based
ArabChat with classification [19]	Arabic MSA	Text	Text	Long	Open	Retrieval-Based
Mobile ArabChat [20]	Arabic MSA	Text	Text	Long	Close (for students of Applied Science University)	Retrieval-Based
Abdullah CITS [4]	Classical Arabic/MSA	Text	Text	Long	Close (teach Islam for children)	Retrieval-Based
LANA CITS [21]	English/classical Arabic	Text	Text	Long	Close (for children with Autism Spectrum Disorder)	Retrieval-Based

B. Speech Conversation Chatbot

The interaction in the speech conversation chatbot is based on the voice as an input, or output, or both. Also, the textual interaction for input or output in this type of chatbot is supported as well with the voice interaction. However, the research on speech conversation chatbots is limited in Arabic. This section presents two related works, that are classified based on the used approach into AIML and pattern matching.

1) *AIML Approach*: Hala [22] is a female robot receptionist located at Carnegie Mellon University in Qatar (CMU-Qatar). Hala accepts and speaks English and Arabic. There are three possible input modes the English, MSA or 'Arabizi' which is Arabic written in English letters. The users interact with the chatbot through the keyboard. Hala responses by producing a voice reply and a text appears next to her face on the screen. The response language depends on the user input language. Hala provides information about campus directions, weather, local events and answer queries regarding her personal life, in an open domain and long conversation style. The conversation between Hala and the user takes an equal number of turns. When the user leaves, Hala will detect the conversation was ended after a defined timeout. The purpose of implementing Hala project was to explore culture of the human-robot interaction in the CMU-Qatar by studying the dialogue patterns such as robot's attributes, covered knowledge bases, and cultural variation in the community of users.

2) *Pattern Matching Approach*: IbnSina [23], [24] is a multilingual conversational robot, that supports Arabic MSA and English. The user interacts with it through text or voice inputs. IbnSina robot responses with audio output, where the language of the response matches the user input language. IbnSina robot generates human interaction dialogue by accessing the online Wikipedia and the stored Quran database. This makes the Chabot of IbnSina robot covering a wide area of topics. Because of that, it replies to general questions, translating words, or answering the question by giving online information, or from the stored books in its database. Also, it gives the user feedback when there are missing information or incorrect spelling. That makes the conversation style of IbnSina open and long.

However, it does not generate new responses, as it depends on the information predefined in the dataset to respond with an appropriate output. The IbnSina conversation system is designed based on object-oriented classes as Wikipedia class and Quran class, that allows the robot to reply with the expected response such as chatterbot class, that enables making a simple conversation and reply to user inquiries. Also, there is a developed chatterbot module, that replies to user inputs.

There are two modules supported in the conversation. First, text-to-speech, where the input is text, and the output is audio. Second, speech-to-text, the input is an audio that is converted to text then processed as the first module to get the speech output. In addition, IbnSina robot supports other features such as read-aloud-text by reading a text image through the camera located in the robot eyes area. The robot interacts with the user by body interactions such as real-time lip syncing, eye blinking, face movement, facial expressions, and shaking hands. Table III shows a summary of the discussed related work for both AIML and pattern matching approach of the speech Arabic chatbots.

From the reviewed studies, we notice that all presented work on Arabic chatbot applications is employing retrieval-based model. That is, the chatbot responses are based on the data pool from AIML files, database, or web pages. Which can limit the capability and usability of the chatbot. Also, we notice that all related work relay on AIML or pattern matching approaches. That may lead to 1) a small size of chatbot's dataset and a restriction to closed domain 2) limits chatbot's response to the user where it requires that the user input matches the chatbot dataset to get the correct response. Moreover, the complexity of Arabic grammar and the user's spelling and grammar mistakes could be one of the reasons for the shortage in capabilities of Arabic chatbots in the literature. Which can explain the limited number of the text and speech Arabic chatbot applications.

TABLE III. A SUMMARY OF SPEECH ARABIC CHATBOTS

Chat bot	Language	Interaction Type		Approach	Conversation Length	Domain	Dataset Model
		Input	Output				
Hala [22]	English / Arabic (Arabizi / MSA)	Text	Text / Voice	AIML	Long	Open	Retrieval-Based
IbnSina [23], [24]	English / Arabic (MSA)	Text / Voice	Voice	Pattern Matching [25]	Long	Open	Retrieval-Based

V. CONCLUSIONS

This paper presents a survey on Arabic chatbots covering twelve different Arabic chatbot studies. They are classified based on the chatbot conversation interaction type into two groups, text and speech conversational chatbots. The studies were presented and evaluated based on the implementation technique, the conversation length and domain, and the model used for the chatbot dataset. The evaluation shows that all the reviewed chatbots were incorporating retrieval-based dataset model. This rises a flag to focus on studying and formalizing Arabic NLP for the conversational agent research. Linguistic complexities hindering Arabic NLP such as morphological ambiguities which means that the word has many meanings, and syntactic ambiguities which means that the sentence has more than one structure, along with the diversity of Arabic dialects are open research challenges that needs cooperation of linguists and computer scientists. Hence, until now, pattern matching and AIML are the ways used to build Arabic conversational agents. Additionally, generative-based and deep learning models are challenging to achieve in Arabic, for the lack of available resources to train the learning model compared to the pool of resources available in English for example. Moreover, there is a shortage in the published

research on Arabic chatbot compared to the available commercial applications that are more sophisticated and developed than the chatbots available in the literature.

REFERENCES

- [1] Z. X. Bingquan Liu, B. W. Chengjie Sun, D. F. W. Xiaolong Wang, and Min Zhang, "Content-Oriented User Modeling for Personalized Response Ranking in Chatbots," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. PP, no. 99.
- [2] B. A. Shawar, "A Chatbot as a Natural Web Interface to Arabic Web QA," *Int. J. Emerg. Technol. Learn. IJET*, vol. 6, no. 1, pp. 37–43, Mar. 2011.
- [3] D. A. Ali and N. Habash, "Botta: An Arabic Dialect Chatbot.," in *COLING (Demos)*, 2016, pp. 208–212.
- [4] O. Alobaidi, K. Crockett, J. O'Shea, and T. Jarad, "Abdullah: An intelligent arabic conversational tutoring system for modern islamic education," in *Proceedings of the World Congress on Engineering*, 2013, vol. 2.
- [5] G. F. L. Chayan Chakrabarti, "A Semantic Architecture For Artificial Conversations," presented at the The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems.
- [6] B. Shawar and E. Atwell, "Accessing an information system by chatting," *Nat. Lang. Process. Inf. Syst.*, pp. 570–584, 2004.
- [7] R. Wallace, "The Anatomy of A.L.I.C.E.," in *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, R. Epstein, G. Roberts, and G. Beber, Eds. Dordrecht: Springer Netherlands, 2009, pp. 181–210.
- [8] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A Survey of Design Techniques for Conversational Agents," in *Information, Communication and Computing Technology*, 2017, pp. 336–350.
- [9] L. Bradeško and D. Mladenčić, "A survey of chatbot systems through a loebner prize competition," in *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, 2012, pp. 34–37.
- [10] K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," 2014, pp. 1724–1734.
- [11] H. N. Io and C. B. Lee, "Chatbots and conversational agents: A bibliometric analysis," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2017, pp. 215–219.
- [12] S. AbdulKader and J. Woods, "Survey on Chatbot Design Techniques in Speech Conversation Systems," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 6, no. 7, 2015.
- [13] B. Shawar and E. Atwell, "An Arabic chatbot giving answers from he Qur'an," in *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*, 2004, vol. 2, pp. 197–202.
- [14] R. Wallace, *The Elements of AIML Style*. Alice AI Foundation, 2003.
- [15] W. Brini, M. Ellouze, S. Mesfar, and L. Belguith, "An Arabic question-answering system for factoid questions," in *2009 International Conference on Natural Language Processing and Knowledge Engineering*, 2009, pp. 1–7.
- [16] Pandorabots, Rosie: Base content for AIML 2.0 chatbot. pandorabots, 2018.
- [17] M. Hijjawi, Z. Bandar, K. Crockett, and D. Mclean, "ArabChat: An Arabic Conversational Agent," in *2014 6th International Conference on Computer Science and Information Technology (CSIT)*, 2014, pp. 227–237.
- [18] M. Hijjawi, Z. Bandar, and K. Crockett, "The Enhanced Arabchat: An Arabic Conversational Agent," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 2, 2016.
- [19] M. Hijjawi, Z. Bandar, and K. Crockett, "User's utterance classification using machine learning for Arabic Conversational Agents," in *2013 5th International Conference on Computer Science and Information Technology*, 2013, pp. 223–232.
- [20] M. Hijjawi, H. Qattous, and O. Alsheiksalem, "Mobile Arabchat: An Arabic Mobile-Based Conversational Agent," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 6, no. 10, 2015.
- [21] S. Aljameel, J. O'Shea, K. Crockett, A. Latham, and M. Kaleem, "Development of an Arabic Conversational Intelligent Tutoring System for Education of children with ASD," in *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2017, pp. 24–29.
- [22] M. Makatchev et al., "Dialogue patterns of an Arabic robot receptionist," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 167–168.
- [23] N. Mavridis, A. AIDhaheeri, L. AIDhaheeri, M. Khanii, and N. AIDarmaki, "Transforming IbnSina into an advanced multilingual interactive android robot," in *2011 IEEE GCC Conference and Exhibition (GCC)*, 2011, pp. 120–123.
- [24] L. Riek et al., "Ibn sina steps out: Exploring arabic attitudes toward humanoid robots," in *Proceedings of the 2nd international symposium on new frontiers in human–robot interaction, AISB, Leicester*, 2010, vol. 1.
- [25] N. Mavridis, "About the Chatterbot Used in IbnSina Conversational Robot. Personal E-mail," 23-Feb-2018.

Learner Cognitive Behavior and Influencing Factors in Web-based Learning Environment

Kalla Madhusudhana

Professor, Department of CSE
CVR College of Engineering, Hyderabad, India

Abstract—In educational institutions, to improve student learning outcome and performance, the information and communication technology has enabled us to embark web-based learning approaches. The traditional web-based learning environment in higher education is aimed at fulfilling the users for most of their deserved learning contents as per the course curriculum. But, in modeling the course curriculum and content, the motivational factors have been left out, through which the learner's cognitive skills development can take place. Therefore, in e-learning courses, this issue needs to be addressed. It can be resolved through subsuming suitable learning objectives and appropriate skills based interactive learning resources, which can enhance thinking skills and cognitive behavior of learner. This paper provides theoretical framework on the pedagogical factors that can influence the quality of students' learning experience and cognitive learning skills in web based learning environment. Furthermore, this study discusses about the role of prior knowledge and learner's thought process model in cognitive based learning environments.

Keywords—*Learning environment; cognitive behavior; influencing factors; pedagogical; knowledge; curriculum content*

I. INTRODUCTION

On Web-based Education Technology, most of the existing research studied about how to design better system architecture to deliver learning resources as per the learner context and interest. To the best of our knowledge, no research exists that tackles the motivational factors that influence Cognitive Learning Behavior (CLB) and the inter relation between Student Cognitive Behavior and Curriculum Content in web-based learning environment. Hwang [1] stated that in developing web-based learning environments there are several issues to be considered; including the ways to stimulate students' interaction and associated pedagogy.

Without employing the appropriate pedagogical and instructional strategies, the expectations of learning outcomes will not be reached [2]. The successful implementation of web based learning that support learner's cognitive skills, need to consider the necessary features such as learners' understandability, performance and behavior.

According to Bloom [3], each level of cognitive development depends upon the learners' behavior and prior knowledge. Learner's prior knowledge facilitates the processing of new incoming information [4] and helps to understand new concepts through the use of real world experience. As stated in [5] the content presentation approach of learning environment could greatly influence student's

navigation behavior towards investigative and cognitive learning.

In higher education, the learner's expectation is to gain not only knowledge in the concerned subject, but also higher-order thinking skills that constitute them as professionals in the domain [6]. As per the cognitive load theory the incorporation of instructional methods in lesson design, can improve learning skills by managing the learner's cognitive load [7]. Managing the learner's cognitive load can help the learner to aware their own thinking and learning process and to select the most suitable way to reach the learning objectives [8]. As per the Bada SO [9], in student-centered learning the attention should be given to the learning environment, as it contribute to the development students' learning needs.

Making the students to actively engage in learning process is more important to academic success [10]. The learner's cognitive capabilities and learning strategies help to prepare learner successful, because they control the learning process to be self-regulative [11]. The process of learning is an active approach, where the learner tries to subsume their current interpretations with past knowledge [12] and make an effort to construct new ideas or decisions.

Learners have different patterns of abilities, strategies and learning styles that are functions of the learner's interaction. The learner needs to know why they need to learn something and use learning strategies to support their strengths. The sequence of learner interactions with the material recognizes the cognitive patterns of each student's learning style [13]. The types of learning resources that are believed to enhance intrinsic motivation need to be integrated into the curriculum contents of the web-based learning environment. For this purpose, the strategies such as case studies, role playing, simulations, and self-evaluation are most useful.

This article starts by defining the learning environment and role of learner's prior knowledge in developing higher level cognitive skills and then we proposed the learners' thought process model in cognitive based learning environments.

II. INFLUENCING LEARNER'S COGNITIVE BEHAVIOR

With the increasing popularity of e-learning programs, educational stakeholders are attempting to promote factors affecting the critical thinking and cognitive capabilities in the virtual education system [14]. Research in educational psychology has provided many principles for encouraging the cognitive processes of learning [15]. The existing literature indicates there is a considerable interest among researchers for

identifying how to incorporate critical thinking support in e-learning environment.

TABLE I. INFLUENCING FACTORS OF CRITICAL THINKING AND COGNITIVE CAPABILITIES OF LEARNER

S.No	Influencing Factors	Influence on Students' learning quality
1	Learner's Prior Knowledge	<p>Helps in improving self – regulated independent learning skills.</p> <p>Provide foundation to skill development, and to set future goals.</p> <p>Influence in the process of understanding the concepts.</p> <p>It should be taken into consideration in instructional design and curriculum planning.</p>
2	Learning Environment	<p>Can pave the way to students to learn independently</p> <p>Supporting student s in self - regulation,</p> <p>Helps in monitoring and adjusting their actions and plans of learning</p> <p>make provision to improve students' learning outcomes,</p>
3	Pedagogical Learning Contents	<p>Make the student in becoming independent and responsible to their learning process.</p> <p>Facilitate better understanding, improved academic performance and improve cognitive skills of the learner.</p> <p>Promote deeper conceptual understanding and investigative behavior of learner.</p>

In web based learning environments, learners are overwhelmed by the increasing amount of information and data from heterogeneous information sources available to them. Furthermore, in the traditional course based e-learning environments the course contents are presented from a specific viewpoint and offering standard content to each individual student [16]. Therefore in this context, we aim at examining

some of the important factors such as the role of learner's prior knowledge, learning environment and Pedagogical factors that influence skilled learning process in web-based learning environment as shown in Table I.

III. ROLE OF PRIOR KNOWLEDGE

Cognitive skills are used to process new stimulus along with recalled knowledge bank where the information has already processed guides learner's behavior [4]. In learning environment the navigational activities of user are influenced by Learning Environment and prior knowledge. The prior knowledge from the previous courses significantly influences student in both knowledge acquisition and the capacity to apply higher-order skills [17]. In student-centered learning the prior knowledge and experience will uniquely influence on individuals knowledge-seeking activity and cognitive repertoire [18].

Prior knowledge has long been considered the most important factor influencing learning and student achievement [19], but not all types of prior knowledge have similar relevance to student achievement. Students need to acquire the important prior knowledge and skills needed when they try to learn more advanced courses in their curriculum. As shown in Fig. 1, the prior knowledge can be defined as a multiple types and various categories of prior knowledge / background knowledge that helps the learner in the mental activity (Cognitive Skills) in the process of Learning.

IV. LEARNING ENVIRONMENT AND PEDAGOGICAL FACTORS

In e-Learning environment student attention be focused on elements that are relevant to understanding and target solution that is based on their skills [7] The main requirement of the web based learning environment is to provide support to the student interactions within a specific cognitive space. In order to encourage higher-order thinking skills of a learner, the learning environment needs to support the cognition-oriented metadata attributes such as Bloom's taxonomy [3]. As per He, Daqing [20], in accessing the learning resources the cognition-oriented metadata attributes helps the learners having different viewpoints, cognitive skills, and knowledge levels of learning materials.

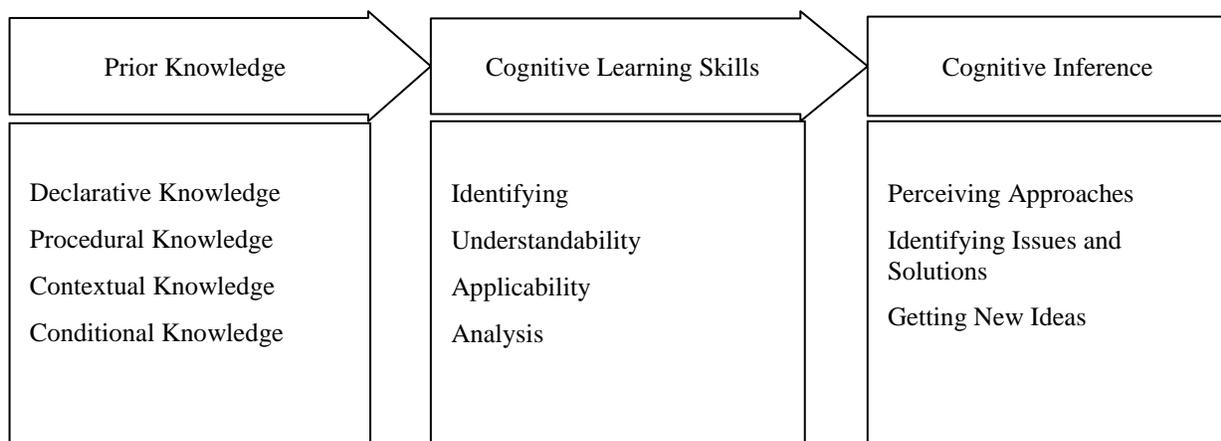


Fig. 1. Various Categories of Prior Knowledge, Cognitive Learning Skills and Inference.

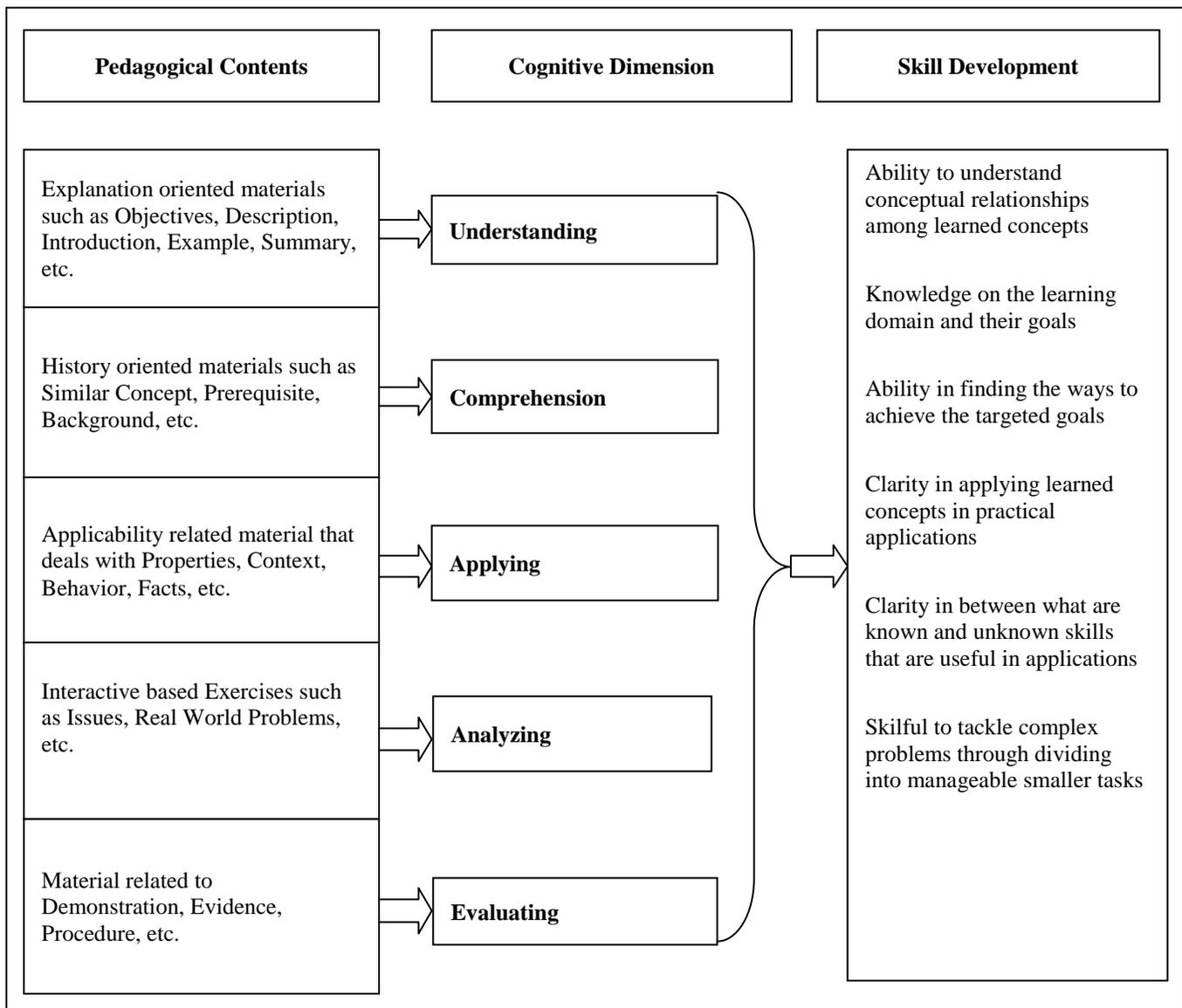


Fig. 2. Pedagogical Metadata Mapped with Cognitive Dimensions and Impact on Learning.

The term critical thinking and learning style are another key factors that deal with the learner’s activity (i.e. Understanding, Applying, and Analyzing), but the learner’s activity can be influenced by means of the content presentation environment and self-regulated learning strategies. The content presentation approach can help in retrieving additional topics which are related contextually and pedagogically with search topic. The self-regulated learning strategies depends on the Procedural knowledge of learner and Interface design (Presentation).

The web-based courseware development is the process of composing a sequence of learning resources that are adapted to particular educational context and scenario. As per the course curriculum, the course designer put together the course contents, where content repository is associated with meta-information.

The knowledge-representation process such as ontological approach for semantic based presentation of the course

contents can improve the understandability of e-learner and the learner is able to refer wide range of different concepts that are related to learning topic. The metadata that helps in characterizing the learning resources as per the pedagogical needs of learner is an important step towards successful e-learning environment. The cognitive dimension of metadata helps in delivering disparate educational resources with various inter-relationships that provide instructional and guided paths as per the learning skills of e-learner.

The learner’s critical thinking and problem-solving skills are depending on various types of cognitive dimensions. The cognitive skills development is the construction of thought processes through exploring various types of learning resources. Fig. 2 shows the cognitive dimensions mapping with pedagogical metadata and impact on the learner’s critical thinking skills. This figure helps us to understand the types of course content resources and metadata model to enhance learner’s cognitive skills.

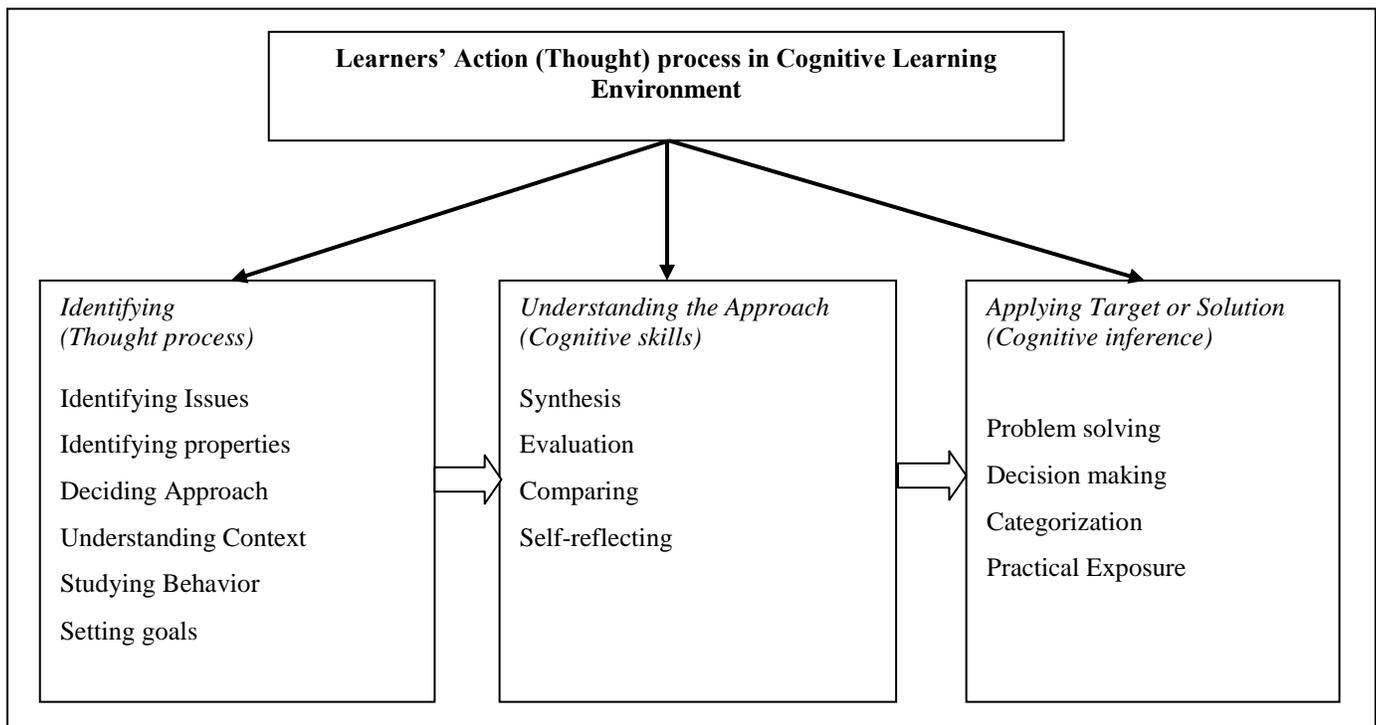


Fig. 3. Learner's Action Process in Cognitive Learning Environment.

V. LEARNERS' THOUGHT PROCESS IN COGNITIVE LEARNING ENVIRONMENTS

The cognitive learning process involves the formation of associations between stimulus and responses [15]. In e-Learning environment the critical thinking is a self-regulatory judgment which results in Identifying, Understanding, and Applying based on the evidential, conceptual, contextual and behavioral considerations upon which the judgment is based. As show in Fig. 3, the learner's thought process (cognitive behavior) model we have presented here focuses on three aspects such as: Identifying, Understanding and Applying. The other way to look at the cognitive skills is to use Bloom's taxonomy.

In the cognitive learning process identifying the problem or issues along with the related learning material with various types of relationships such as context, application, behavior, etc. is the learner's basic activity.

Understanding is the process of reformulate learning topic with known concepts, and trying to predict about what kind of information and strategies need to know so as to get complete awareness on the topic.

Applying knowledge and learned procedures to realistic problems through integrating skills such as categorization, decision making, practical exposure etc. would be an important part of the cognitive learner behavior.

Problem-centered activities help the learner to experience the correct solution that requires integration of knowledge and skills. The Interaction oriented material can provide the foundation for problem-cantered activities. The integrated practical exercises test the extent to which students can apply knowledge.

Decision making can be considered as a process of deciding about what to do, through Evaluation or Comparing with reference to a standard "benchmark". Learning and decision making are inextricably linked and decision-making capability improves student learning standard.

Categorization is a fundamental ability that allows the learner to organize his knowledge in a hierarchy according to complexity and react appropriately for further assessment, useful predictions and to identify prerequisites. Categorizing and extracting relationships are the two most familiar thinking skills.

The integrated practical exercises test the extent to which students can apply knowledge and learned procedures to realistic problems.

VI. CONCLUSION

In this paper, we discussed the integration of course curriculum with factors that influences learning skills, will mobilize the learner's cognitive behavior and learning outcome significantly. The e-learning courses need to be developed with suitable learning objectives and appropriate interactive learning resources. Learner's capability or thought process related to knowledge and comprehension, can have a positive impact on various capabilities of the learner and leads to cognitive skills such as synthesis, evaluation etc. One can group the skills and become autonomous for their decisions according to target action. Developing higher level cognitive skills can encourage the learners towards creativity, solving problems, practical exposure etc. For this, we need to identify the kinds of learning topics that would be included in a curriculum that would mobilize the learner towards understandability and applicability of learned concepts.

To transform the existing formal education into new way of learning and to propose an ideal framework for the standardization of the Cognitive Learning Environment, we are currently designing and developing the system prototype which will be implemented and evaluated soon.

REFERENCES

- [1] Hwang, Wu-Yuin, Chin-Yu Wang, Gwo-Jen Hwang, Yueh-Min Huang, and Susan Huang. "A web-based programming learning environment to support cognitive development." *Interacting with Computers* 20, no. 6 (2008): 524-534.
- [2] Eison, Jim. "Using active learning instructional strategies to create excitement and enhance learning." *Jurnal Pendidikantentang Strategi Pembelajaran Aktif (Active Learning) Books* 2, no. 1 (2010): 1-10.
- [3] Bloom, B.S. and Krathwohl, D.R. *Taxonomy of Educational Objectives: The Classification of Educational Goals Handbook I: Cognitive Domain*, Longmans, Green, NY; (1956).
- [4] Brod, Garvin, Markus Werkle-Bergner, and Yee Lee Shing. "The influence of prior knowledge on memory: a developmental cognitive neuroscience perspective." *Frontiers in behavioral neuroscience* 7 (2013): 139.
- [5] Madhusudhana, Kalla. "The Cognitive Dimension and Course Content Modeling: An Ontological Approach." *International Journal of Emerging Technologies in Learning (IJET)* 12, no. 05 (2017): 181-188.
- [6] Petchtone, Puangtong, and Sumalee Chaijaroen. "The development of web-based learning environments model to enhance cognitive skills and critical thinking for undergraduate students." *Procedia-Social and Behavioral Sciences* 46 (2012): 5900-5904.
- [7] Clark, Ruth, and Gary L. Harrelson. "Designing instruction that supports cognitive learning processes." *Journal of athletic training* 37, no. 4 suppl (2002): S-152.
- [8] Demirel, Melek, İlkay Aşkın, and Esed Yağcı. "An investigation of teacher candidates' metacognitive skills." *Procedia-Social and Behavioral Sciences* 174 (2015): 1521-1528.
- [9] Bada, Steven Olusegun, and Steve Olusegun. "Constructivism learning theory: A paradigm for teaching and learning." *Journal of Research & Method in Education* 5, no. 6 (2015): 66-70.
- [10] Gettinger, Maribeth, and Jill K. Seibert. "Best practices in increasing academic learning time." *Best practices in school psychology IV* 1 (2002): 773-787.
- [11] Zumbunn, Sharon, Joseph Tadlock, and Elizabeth Danielle Roberts. "Encouraging self-regulated learning in the classroom: A review of the literature." *Metropolitan Educational Research Consortium (MERC)* (2011): 1-28.
- [12] Kay, Denise, and Jonathan Kibble. "Learning theories 101: application to everyday teaching and scholarship." *Advances in physiology education* 40, no. 1 (2016): 17-25.
- [13] Romanelli, Frank, Eleanora Bird, and Melody Ryan. "Learning styles: a review of theory, application, and best practices." *American journal of pharmaceutical education* 73, no. 1 (2009): 9.
- [14] Gharib, Mitra, Mitra Zolfaghari, Rita Mojtahedzadeh, Aeen Mohammadi, and Atoosa Gharib. "Promotion of critical thinking in e-learning: a qualitative study on the experiences of instructors and students." *Advances in medical education and practice* 7 (2016): 271.
- [15] Hinojosa, Luz Marina Méndez. "Contributions of Educational Psychology to University Education." *Psychology* 6, no. 03 (2015): 177.
- [16] Souto, Maria AM, Mariano Nicolao, Rosa M. Viccari, José PM de Oliveira, Regina Verdin, Karine Beschoren, Milton Madeira, and Renata Zanella. "Web-adaptive training system based on cognitive student style." In *TeLE-Learning*, pp. 281-288. Springer, Boston, MA, 2002.
- [17] Hailikari, Telle, Nina Katajavuori, and Sari Lindblom-Ylänne. "The relevance of prior knowledge in learning and instructional design." *American Journal of Pharmaceutical Education* 72, no. 5 (2008): 113.
- [18] Hannafin, Michael J., and Kathleen M. Hannafin. "Cognition and student-centered, web-based learning: Issues and implications for research and theory." In *Learning and instruction in the digital age*, pp. 11-23. Springer US, 2010.
- [19] Dochy FJRC. *Assessment of Prior Knowledge as a Determinant for Future Learning: The use of prior knowledge state tests and knowledge profiles*. Utrecht/London: Lemma BV; 1992. pp. 43-72.
- [20] He, Daqing, Ming Mao, and YefeiPeng. "DiLight: a Digital Library based E-Learning Environment for Learning Digital Libraries." In *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, vol. 2006, no. 1, pp. 2845-2852. 2006.

New Hybrid Task Scheduling Algorithm with Fuzzy Logic Controller in Grid Computing

Younes Hajoui, Omar Bouattane, Mohamed Youssefi, Elhocein Illoussamen

Laboratory SSDIA

ENSET Mohammedia, Hassan II University of Casablanca Mohammedia 28999, Morocco

Abstract—Distributed heterogeneous architecture is extensively applied to a diversity of large scale research projects conducive to solve complex computational problems. Mentioned distributed systems consist of multiple heterogeneous linked processing units used to handle the continuous arrival jobs. The tasks scheduling problem is concerned with resource allocation strategies to assign jobs to available computing resources. The load balancing of linked resources becomes a main issue to select in each task schedule the adequate computing resource. Our proposal consists of combining Q-learning with ACO (Ant Colony Optimization) to solve the tasks allocation dilemma. In our proposed Fuzzy Hybrid Framework, Fuzzy ants are used to calculate at each scheduling operation, the novel reward values whereas Q-learning is used to select the suitable Worker Machine. The simulation findings confirmed the efficiency of the proposed framework due to the significant decrease of the makespan.

Keywords—Distributed systems; computational problems; load balancing; Q-learning; ACO; fuzzy hybrid framework

I. INTRODUCTION

The tasks scheduling problem is concerned with resource allocation strategies to assign jobs to available computing resources. The load balancing of linked resources becomes a main issue to select in each task schedule the adequate computing resource.

Due to the heterogeneity of arrival tasks and uneven nodes performance, some nodes work more than others. Therefore, to achieve equal distribution and optimal use of resources, scheduling need to be fair, well studied and strategic [1,2].

In [27], the authors propose a global taxonomy which is used to classify frequently encountered types of job scheduling, facilitate researchers to build on prior art, increase new research visibility, and minimize redundant effort.

In the literature, load balancing algorithms can be classified into centralized, decentralized or hierarchical categories [3].

In centralized scheme, tasks are scheduled first to a central resource then this central node decides how to assign received tasks to executors. A major disadvantage of using central node is that it must not fail because it should ensure the allocation of tasks.

Decentralized scheme does not contain a central scheduler, scheduling decisions is done by all resources in the distributed system [4, 9, 11, 29]. This model suffers from several

weaknesses and especially the cost resulting from the involvement of all resources in the balancing procedure.

In hierarchical model [30], the responsible schedulers are ordered in a hierarchy. This model results from the hybridization of the centralized and decentralized model. Each scheduler is responsible on the schedulers which are below at lower levels and is under the orders of the schedulers from above at higher levels.

Further, load balancing algorithms can be classified into three categories: static, dynamic or adaptive algorithms.

The approach in the static balancing system assigns the tasks by unique and definitive allocation, to the processors or nodes in parallel architectures [5], [6], [28], [30]. Furthermore, the static algorithms don't have the ability to deal with the dynamic changes of such environments. This problem especially arises in distributed systems, where some external variables such as network load and waiting for results of other tasks, make difficult the effective scheduling of tasks. Also, the continuous arrival of new tasks makes the scheduling difficult by a static load balancing approach. In dynamic environments, it is even possible that a static balancing creates major imbalances greater than the balancing produced by a random distribution of tasks. So, the need to adapt the initial machine performance estimation is justified. Dynamic load balancing approach considers, for task scheduling, the current processor load [7, 9,12].

Recently, many load-balancing schemes based on mobile agents have been proposed. The MAS [13] (Mobile Agent Systems) are widely used to offer solutions to dynamic and complex application domains. The main characteristic of these intelligent systems is the migration. The ability of agent migration facilitates the implementation of strong dynamic load balancing strategy. The migrant agent selection is relied on the strategy adopted by the load balancer while the destination is related to the lightly loaded machines.

The migration decision is taken by a centralized load balancer agent that activates the migration process when it is obligatory. The centralized control is not suitable for a dynamic scheme since it must collect data more regularly than the non-centralized one, leading to the overload of the network traffic [14].

In [15], a centralized load balancing scheme is proposed. The principal measure for selecting a node is constructed on job's execution time, while location rule is constructed on cooperation with cluster nodes. A special agent in each node is

in charge for gathering the occupancy rate and the local resource usage quantity. The migration choice is founded on the comparison given to an assumed load threshold value.

In [16] authors suggest a new framework for job scheduling founded on mobile agents. Their proposed model uses a dispatcher agent to schedule parallel jobs to worker agents. Each worker agent is installed in a node of the distributed system giving to a load balancing strategy. A test of application, associated to the distributed image processing, was presented to judge the performance of the framework. Additional work in [17] used a mobile agent, to migrate the jobs from overloaded nodes to the under loaded ones. In the used distributed system, each job should be allocated to a VPU (virtual processing unit). The VPUs connected with each other asynchronously by exchanging through their ports ACL messages (FIPA-ACL). Exchanged messages contain data and jobs to be performed.

In [18,19] the authors centered their research on studying load balancing necessities in a distributed system and planned a design and implementation of an advanced load balancing scheme for grid environment via machine learning. Their method is equilibria to the load dynamically. It uses initial load data kept in the database at the primary level of the procedure. Once a load imbalance arises, the recent load data is collected and warehoused as raw data. Later, numerous machine-learning algorithms have been used to process and investigate the logged data. As a final step, the rules are automatically engendered by data mining methods and used for migrating jobs to rebalance loads.

Recently Multi-agent learning methods have been extensively used in the problematic of resource allocation in the Grid. In [20] the authors present Reinforcement learning in which the agents learn through a trial and error to familiarize to all variations such as the changing resource capacities, latencies, or resource failure, by getting rewards for its actions. The Agents give a score rewarding each machine based on its role to reduce the maximum completion time (makespan).

The tasks scheduling has been proved as a NP-hard problem accordingly [21, 31, 32]. Hence, the use of swarm intelligence systems has become very suitable to deal with the difficulty of such problems [22]. Ant colony optimization is one of the well-known meta-heuristics that is largely used in both path finding and load balancing [23, 22]. In [23] Authors suggest two new distributed swarm intelligence inspired load-balancing algorithms. The first and the second algorithm are correspondingly based on ant colony and on particle swarm optimization. The test of their proposed model is conducted by means of GridSim, which is a platform of simulation based on Java [24]. The robustness of their two strategies is assessed using performance criteria such as makespan and load balancing ratio.

In [25], the authors suggest a new scheme inspired load-balancing algorithms founded on the use of ant colony optimization. In the setting of their exploration, the load balancer is used as an ant which selects, for the recent job, the worker machine having the higher amount of pheromone.

Recently Multi-agent learning for load balancing problems has been extensively treated in the literature. In [26] the authors present machine learning in which the agents learn through the previous experiments completed by the scheduler. It is through test and mistake that the agent learns and progresses his tactic. The Agents allocate a score rewarding each worker machine based on its performance in the past. The principal goal of these teams of cooperative agents is maximizing the global reward, which will later reduce the overall execution time (makespan).

In this paper, we propose a new Framework for task scheduling based on hybridization of Q-Learning and ant-colony optimization technique. Ants are used to calculate reward and Q-Learning is used to schedule the current task to the appropriate worker. In the planned model, a grid manager agent is involved to allocate received jobs to the available worker agents according to the precise decisions to minimize the total execution time (makespan). The proposed framework is constructed by means of three layers, which are the user task producer layer, the scheduling load balancing layer and the workers layer. The implementation of the proposed method uses the agents based middleware for distributed programming JADE tool [8].

The structure of this paper is as follows: In Section II, we formulate and describe the problem presented in this work. Next in Section III we present the technical backgrounds used to develop the proposed scheme. In Section IV, we present the load balancing system used in task routing. In Section V, an example of application using Multiple Program Multiple Data (MPMD) architecture for the distributed image processing, is presented to assess the performance of the proposed framework. In the last section, a conclusion and perspectives are presented.

II. PROBLEM DESCRIPTION

Basic assumptions and notations used in this paper are listed below in Table I:

In our study, the total execution time: makespan can be expressed as follows:

$$\theta P(t) = \underset{i=1}{\overset{N}{\text{Max}}}(\theta P_i(t)) \quad (2)$$

Where:

$$\theta P_i(t) = \left[\left(\frac{\sum_k \tau_k}{\tau_0} \right) * L_i + TE_i(t) \right] / NBC_i \quad (3)$$

k : Is the index of task T_k listed on the queue of the node M_i at time t.

$TE_i(t)$ can be formulated as follows:

$$TE_i(t) = \sum_{i=1}^{|Q_i|} \tau_i \quad (4)$$

TABLE I. BASIC ASSUMPTIONS AND NOTATIONS

Notation	Meaning
$[N] = \{1, 2, \dots, n\}$	Array of available resource workers.
$[T] = \{1, 2, \dots, m\}$	Array of Tasks to be executed.
L_i	Speed of the network linking the node M_i with the Dispatcher.(see section V. A for further details) $L_i = \theta_i - \theta PN_i$ (1)
P_i	Computational power of the worker machine M_i .
C_i	Complexity of task $T_i \in [T]$.
τ_i	Estimation execution time of $T_i \in [T]$.
$TE_i(t)$	Estimated times of all Tasks wait in line on node (M_i) at time t .
$ Q_i $	Number of the wait in line tasks in machine M_i .
NBC_i	Number of cores of node M_i (CPUs).
$NbT_i(t)$	Number of the wait in line tasks on node M_i at time t .
$r_i(t)$	Reward of machine i at time t .
$s_i(t)$	State of the machine (M_i) at time t . $s_i(t) = \{L_i, P_i, NBC_i, \tau_i, TE_i(t), NbT_i(t), r_i(t)\}$.
$S(t)$	$S(t) = \{s_1(t), s_2(t), \dots, s_N(t)\}$.
$R(t)$	$R(t) = \{r_i(t)\}_{i=1..N}$.

Finally, the scheduling problem can be formulated as minimization problem as given below:

$$Min[\theta P(t)] = Min_{\tau \leq t} \left(\underset{i=1}{Max}(\theta P_i(t)) \right) =$$

$$Min_{\tau \leq t} \left[\left[\left(\frac{\sum_k \tau_k}{\tau_0} \right) * L_i + TE_i(t) \right] / NBC_i \right] \quad (5)$$

III. TECHNICAL BACKGROUND

A. Ant Colony Optimization for the Travelling Salesman Problem (TSP)

The TSP is a complex problem widely studied in operations research discipline. The problem consists in finding a shorter path on a map that a traveler can choose to visit a list of cities. Each city must be visited once. At the end of the trip, the traveler must return to the starting city. The TSP problem is solved by Dorigo et al. [20] by using the ant colony optimization method (ACO), which engages a set of artificial ants acting parallel searches on a map. The engaged ants choose the following city based on inter-city distances and the amount of pheromone deposited on the paths connecting the cities. However, the pheromone evaporation is also must be taken into consideration to deviate from the local optimum solution.

For ant K , the probability of traveling from city i to city j is given by the next formula [13, 14].

$$P_{ij}^k = \begin{cases} \frac{(\tau_{ij})^a \cdot (\eta_{ij})^b}{\sum_{l \in J_i^k} (\tau_{il})^a \cdot (\eta_{il})^b} & \text{if } j \in J_i^k \\ 0 & \text{if } j \notin J_i^k \end{cases} \quad (6)$$

Where:

J_i^k : Represent all neighbor cities of i of the k^{th} ant.

τ_{ij} : Represent the value of pheromone on the path linking city i by city j , a and b control the importance of η_{ij} and τ_{ij} .

$$\eta_{ij} = \frac{1}{d_{ij}} \quad (7)$$

Where: d_{ij} : Is the distance between cities i and j .

τ_{ij} is continuously updated between i and j by the following formula:

$$\tau_{ij}(t) = (1 - \rho) \tau_{ij}(t) + \sum_{k=1}^m \Delta \tau_{ij}^k(t) \quad (8)$$

Where:

ρ : Pheromone evaporation coefficient.

$\Delta \tau_{ij}^k$: Deposited Quantity of pheromone by k^{th} ant.

B. Q-learning

Reinforcement learning (RL) is a recent technique to deal with stochastic and complex problems. By the successive experiences and by trial and error exchanges between executers agent and their residing environment, the agents learn how to act optimally and to cope with more complex situations. This Reinforcement learning model consist of an agent observing its environment, choosing an action from the present state and at that moment winning a positive or negative reward to the action selected. The intention of the agent, during its exploration, is to maximize its total upcoming positive reward and avoid as possible any penalties. Q-learning [25, 26] figure among the most known RL algorithms. Following the same RL philosophy, Q-learning seeks to solve optimally any specified Markov decision problem via a Q function. Q-value can be calculated by (9) listed in the algorithm below:

Algorithm 1: Q-learning Algorithm

Begin

Initialize Q (s, a) arbitrarily

Initialize s

Repeat

Select a' From s' using rule resulting from Q;

Take action a , observe r, s' ;

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma(Q(s', a') - Q(s, a))]; \quad (9)$$

$s = s'$;

Until s is final state;

End.

IV. PROPOSED WORK

In this section, we propose the fuzzy hybrid algorithm for solving the problem of scheduling tasks on heterogenous distributed machines. Next, we present a case study and show the efficiency of the proposed hybrid method.

A. Architecture Description

This study purposes to solve a scheduling problem on heterogenous parallel machines. An effective scheduling algorithm plays a significant role in an effective supervision of the grid and so in reducing the maximum completion time. In our suggestion, the load balancing is reached by using a fuzzy hybrid algorithm and a hierarchical mobile agent system.

As shown in Fig. 1, five types of mobile agents are engaged in our proposal: The Producer-Agent, the Tester-Agent, the Dispatcher-Agent, the Controller-Agent and the Worker-Agent. The Producer agent is the one that characterizes the creators of tasks as: web application, mobile application, embedded system (IOT), expert agent (human) and so. The Tester-Agent approximate by previous experiments the execution time of novel tasks. The Dispatcher agent is a type of central manager of the grid and is in charge for distributing new arrival tasks among available workers. Controller-Agent is responsible for continuously monitoring the status of the workers and Worker-Agent performs the tasks received from the dispatcher.

Each machine that seeks to join the grid, at the 3rd layer, as a worker node to participate on the computation must follow these three steps called referencing phase [26]:

- 1) Send a request to the dispatcher to join the workers.
- 2) Receive and perform the referencing task: T0.
- 3) Communicate to the Dispatcher the results of T0 execution: Li, Pi, NBCi.

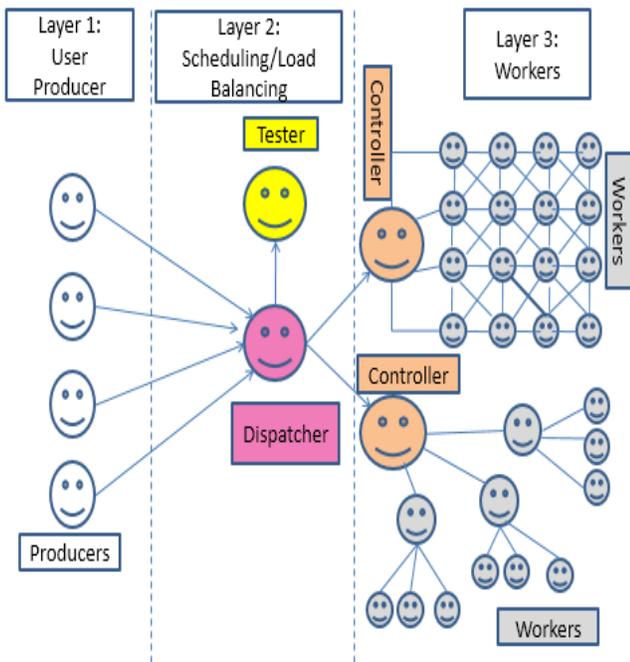


Fig. 1. Framework Architecture [26].

B. Proposed Fuzzy Hybrid Algorithm: FHA

The design of the proposed fuzzy hybrid algorithm is inspired from the combination of Q-Learning and Ant Colony algorithms. In this method, the hybrid algorithm performs in two parallel phases. In the first phase, ant colony algorithm involves a set of artificial ants acting parallel searches on network links between Dispatcher and resources. Each ant A_i receives from the controller machine, belonging on its network link, the state $s_i(t)$ of the worker M_i . All ants $\langle A_i \rangle_{i=1..N}$ are inside the Dispatcher node and calculate $P_{di}(t)$ according to (10).

$$P_{di}(t) = \frac{\left[r_i(t)^\alpha * \frac{1}{NbT_i(t)^\beta} * P_i^\gamma * \frac{1}{L_i^\theta} * NbC_i^\delta * \frac{1}{TE_i(t)^g} \right]}{\sum_{j=1}^N \left[r_j(t)^\alpha * \frac{1}{NbT_j(t)^\beta} * P_j^\gamma * \frac{1}{L_j^\theta} * NbC_j^\delta * \frac{1}{TE_j(t)^g} \right]} \quad (10)$$

Where:

$$\tau_{di} = r_i(t) \quad (11)$$

$$\eta_{di}^b = \frac{1}{NbT_i(t)^\beta} * P_i^\gamma * \frac{1}{L_i^\theta} * NbC_i^\delta * \frac{1}{TE_i(t)^g} \quad (12)$$

$R_i(t)$ refers to the pheromone deposited on the network link connecting the Dispatcher with the worker machine M_i . It's also refers to the reward of the worker machine M_i at time t. The Dispatcher rewards powerful machines by a positive value and discourages weak machines by a negative value. The rewards and penalties are continuously calculated in order to supply the Q-Learning by the updated values to select the appropriate worker.

To measure the worker reward, fuzzy logic process is implemented by following in order of these three steps:

1) *Fuzzification of Input:* The first step in the fuzzy inferencing method is the fuzzification. It consists to convert crisp inputs into fuzzy inputs. Crisp inputs are precise inputs calculated in real time by the Dispatcher. As shown in Fig. 2, $P_{di}(t)$ and $\theta P_i(t-1)$ are the two measured crisp inputs passed into the Fuzzy system for treatment to calculate associated crisp output: reward.

In addition, it is important to note that $\theta P_i(t-1)$ refers to the execution time of each node in the last previous schedule. This parameter shows the performance of each node in the history and is considered important next to $P_{di}(t)$ to calculate the value of the worker's reward.

For each crisp input, a membership function is associated. The two following figures: Fig. 3 and 4 shows the curve of membership functions for : "Node rapidity" and "Node state".

Regardless of the worker historical, a value of $P_{di}(t)$ that tends to 1 indicates that the worker M_i is the most candidate likely to receive the current task. This also shows that M_i is the most under loaded among all the machines in the grid.

TABLE II. IF THEN-RULES

If-Then Rules	
1	if node state is (very overloaded or overloaded) and node rapidity is (slow or normal) then node reward is very bad
2	if node state is normal and node rapidity is (slow or normal) then node reward is bad
3	if node state is normal and node rapidity is speed then node reward is good
4	if node state is underloaded and node rapidity is slow then node reward is bad
5	if node state is underloaded and node rapidity is normal then node reward is good
6	if node state is very underloaded and node rapidity is (slow or normal) then node reward is good
7	if node state is (underloaded or very underloaded) and node rapidity is speed then node reward is very good

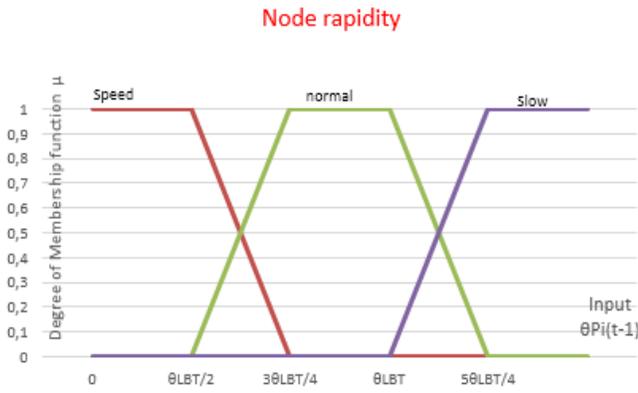


Fig. 2. Input-Node Rapidity.

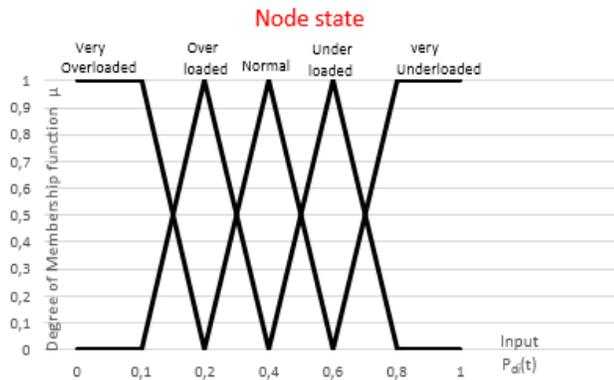


Fig. 3. Input-Node State.

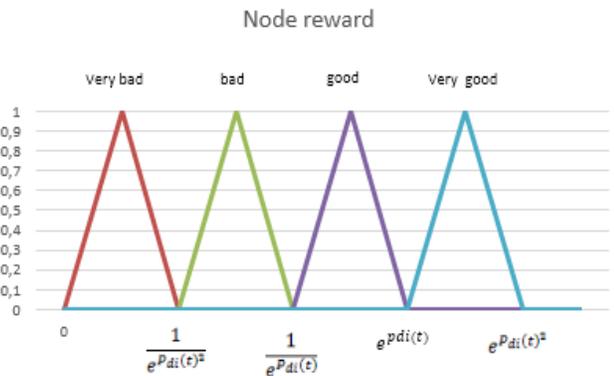


Fig. 4. Output-Node Reward.

The fuzzy system output is a result of all the inputs and the rules. Fig. 4 shows the curve of membership functions for the output: Node reward.

2) *Fuzzy Inference Process-Rules*: Fuzzy inference is the procedure of expressing from specific inputs, an output by using fuzzy logic. This process implies membership Functions, Logical operations, and If-Then Rules. The nature of the rules is the most important parametrization of fuzzy logic systems. It allows concluding the output based on the inputs and the rules. Table II shows the list of fuzzy rules employed to capture the imprecise methods of worker rewarding.

3) *Defuzzification*: Defuzzification is the last step that returns crisp output from the fuzzy sets. There are several types of defuzzification methods. The following are the well-known methods: Center of Sums Method (COS), Center of gravity (COG) / Centroid of Area (COA) Method, Center of Area / Bisector of Area Method (BOA), Weighted Average Method and Maxima Methods.

The COG defuzzification is more commonly used fuzzy mathematics method; it defines the output as corresponding to the abscissa of the center of gravity of the surface resulting from the combination of the conclusions and the rules in order to apply the output found to the original problem.

At each computation of $P_{di}(t)$, the reward is calculated by using COG defuzzification as follows:

$$\text{Reward}^* = \frac{\int x\mu(x)dx}{\int \mu(x)dx}$$

Hence:

$$\tau_{di} = r_i(t) = \frac{\int x\mu(x)dx}{\int \mu(x)dx}$$
(13)

In the inspired ACO algorithm, the smell of pheromone, as shown in (13), has a direct effect on the calculation of future rewards by the ants. Updating pheromone means that even if a machine has a favorable reward, it does not prevent the dispatcher from continuously reviewing its workload in order to assign it a new value proportional to its state among the other resources. In the second phase, Q-Learning is used to search an optimal action-selection strategy for the tasks allocation. Based on the immediate reward $f(r_i(t))$, the agent Dispatcher updates its estimate for Q by its latest observation from the grid environment. It calculates Q_i for each machine M_i according to (14). Then, it selects the machine having the great value of Q_i for the current task.

$$Q_{i+1}(s_i, a_i) = Q_i(s_i, a_i) + \alpha(r_i(t) + \gamma \max_{a'}(Q_i(s', a') - Q_i(s_i, a_i)))$$
(14)

The grid manager executes parallelly both collaborative algorithms to solve the problem described in this work. Ants

are used to calculate reward and Q-Learning is used to schedule the current task to the appropriate worker.

V. PERFORMANCE EVALUATION AND DISCUSSIONS

In this section, the configuration of the used resources and the results obtained from the fuzzy hybrid algorithm test are described and presented.

The entire system is developed in Java and is tested on a cluster of 10 heterogenous workers. The proposed load distribution process purposes to deploy the agent features to develop a self-directed organization by means of a real multi-agent system based on the JADE platform (Java Agent Development Framework).

The associated scheme aims to build a multi-agent system to schedule jobs on a cluster of 10 heterogeneous machines. The used method is based on the intelligent agents, which must be able to delegate specific tasks.

The configurations of the 10 heterogenous workers is determined by executing a referential task T_0 by each machine before joining the grid. The machines have four distinct capacities: P_i, L_i, NBC_i . The scheduler must take into account the heterogeneity of node capacities before making scheduling decisions.

A. Referencing Phase

In this test example, there are ten heterogenous worker nodes that are shown in Table III. These worker configurations, shown below, are determined by executing a referential task T_0 [26].

Note that:

θ_i : Total Time required to perform T_0 .

θPN_i : Total Time required to perform T_0 on Node $N_i(i=1..n)$.

$$L_i = \theta_i - \theta PN_i \tag{15}$$

$$P_i \text{ is inversely proportionate to } \theta PN_i. P_i = 10^3 / \theta PN_i \tag{16}$$

TABLE III. PARAMETERS CALCULATION BY REFERENCING PHASE

Node i	Pi (ms)	Li (ms)	NbCi	Reward:
0	80	2	4	1
1	150	10	2	1
2	120	8	1	1
3	70	10	4	1
4	100	8	8	1
5	90	10	2	1
6	160	3	1	1
7	170	5	16	1
8	120	1	2	1
9	145	3	4	1

B. Q-Learning and ACO Parametrization

The performance metric in searching optimum results depends principally on the parameterization of both Q-Learning and ACO operators.

The best Q-Learning operators found are shown in Table IV.

TABLE IV. BEST PARAMETER SETTINGS OF THE Q-LEARNING OPERATORS

α	γ
0.3	0.4

The best ACO operators found are shown in Table V.

TABLE V. BEST PARAMETER SETTINGS OF THE ACO OPERATORS

α	β	γ	θ	δ
3	3	3	5	2

C. Load Balancing Theoretic (LBT)

Theoretically, the load balancing can be calculated for a giving system S, having at time t, N distributed resources and an overall execution time T, each resource must have a workload of execution time around the theoretical value: $LBT = T / N$ which is impossible experimentally [26].

D. Scheduling Test by using Fuzzy Hybrid Algorithm (FHA)

The test experiments were generated using a set of NbT heterogenous tasks ($NbT = 1000$). To evaluate and measure the system performance, FHA algorithm is put under dissimilar system loads complexity, we select randomly for each task T_i :

$$\begin{cases} \tau_i = 20i & \text{if } i \text{ is an odd number.} \\ \tau_i = 20i^2 & \text{otherwise} \end{cases}$$

Our hypothesis is tested by making a comparative experiment with the results obtained with scheduling by using Ant colony optimization ACO [25] and by Q-Learning QL [26] under the same controlled configurations.

The scheduling findings are as follows:

Table VI shows the distribution results by using ACO, Q-learning, FHA method, whereas Fig. 5 shows a comparison between their curves duration.

TABLE VI. DISTRIBUTION RESULTS BY USING ACO, Q-LEARNING, FHA

Node Ni	θPi (ms) FHA	θPi (ms) Q-L	θPi (ms) ACO
0	70694609	91214279,5	50694609
1	90009970	56255662	40009970
2	40009970	37788468	70171256
3	60171256	34039604,5	90171256
4	60171256	36351290	41397431
5	143171256	42744272	110171256
6	78018383	148449863	58018383
7	129563908	89091486,9	149563908
8	55679447	165171256	65679447
9	85106444	99118657,5	55106444,3

VI. CONCLUSION

In this paper, we have developed a Fuzzy hybrid method called FHA to solve the problem of tasks scheduling. For efficiency purpose, the proposed Framework simultaneously applies two algorithms that solve the same problem. The aim of the proposed hybrid model is to combine the effectiveness of Q-Learning and ACO to reduce the overall execution time and then comes imbalance among available resources. The experiment results showed that the proposed hybrid algorithm has achieved a perfect convergence in terms of the load balancing among available nodes in grid as well as improving the optimal solution.

The proposed method can be extended to hybridize other metaheuristics with RL algorithms in order to minimize as possible the total tardiness.

REFERENCES

- [1] X. Tang, S. Chanson, "Optimizing static job scheduling in a network of heterogeneous computers," the 29 International Conference on Parallel Processing, 2000.
- [2] K. Li. Optimal load distribution in nondedicated heterogeneous cluster and grid computing environments. Journal of Systems Architecture, 54(1):111–123, 2008.
- [3] Kandagatla, C. (2003). Survey and Taxonomy of Grid Resource Management Systems, University of Texas, Austin. [Online] Available: <http://www.cs.utexas.edu/users/browne/cs395f2003/projects/File:KandagatlaReport.pdf>.
- [4] Y. Cao, W. Yu, W. Ren, and G. Chen, "An overview of recent progress in the study of distributed multi-agent coordination," IEEE Trans. Ind. Informatics, vol. 9, no. 1, pp. 427–438, 2013.
- [5] Kameda, H., Li, J., Kim, C., Zhang, Y.: Optimal Load Balancing in Distributed Computer Systems. Springer, London (1997)
- [6] Penmatsa, S., Chronopoulos, A.T.: Price-based useroptimal job allocation scheme for Grid systems. In: Proceedings of 20th IEEE International Parallel and Distributed Processing Symposium, Rhodes, 2006
- [7] Dhakal, S., Hayat, M.M., Pezoa, J.E., Yang, C., Bader, D.A.: Dynamic load balancing in distributed systems in the presence of delays: a regenerationtheory approach. IEEE Trans. Parallel Distrib. Syst. 18(4), 485–497 (2007)
- [8] Dobber, M., Koole, G., Mei, R.: Dynamic load balancing experiments in a Grid. In: Proceedings of IEEE International Symposium on Cluster Computing and the Grid, Cardiff, 2005
- [9] Penmatsa, S., Chronopoulos, A.T.: Dynamic multi-user load balancing in distributed systems. In: Proceedings of 21st IEEE International Parallel and Distributed Processing Symposium, Long Beach, 2007
- [10] Shah, R., Veeravalli, B., Misra, M.: On the design of adaptive and decentralized load balancing algorithms with load estimation for computational Grid environments. IEEE Trans. Parallel Distrib. Syst. 18, 1675–1686 (2007)
- [11] Arora, M., Das, S.K., Biswas, R.: A de-centralized scheduling and load balancing algorithm for heterogeneous Grid environments. In: Proceedings of International Conference on Parallel Processing Workshops, pp. 499–505. IEEE, Piscataway (2002)
- [12] Zheng, Q.: Dynamic load balancing and pricing in grid computing with communication delay. J. Grid Comput. 6, 239–253 (2008)
- [13] F. L. Bellifemine, G.Caire, and D. Greenwood, "Developing Multi Agent Systems with JADE". Wiley, 2007.
- [14] Maha A. Metawei, Salma A. Ghoneim ,Sahar M. Haggag , Salwa M. Nassar .'Load balancing in distributed multi-agent computing systems', Ain Shams Engineering Journal, (), pp. 237–249. (23 May 2012)
- [15] Cho ChoMyint, Khin Mar LarTun, A Framework of Using Mobile Agent to Achieve Efficient Load Balancing in Cluster. In: Proc. 6th Asia Pacific symposium on information and telecommunication technologies; 2005.

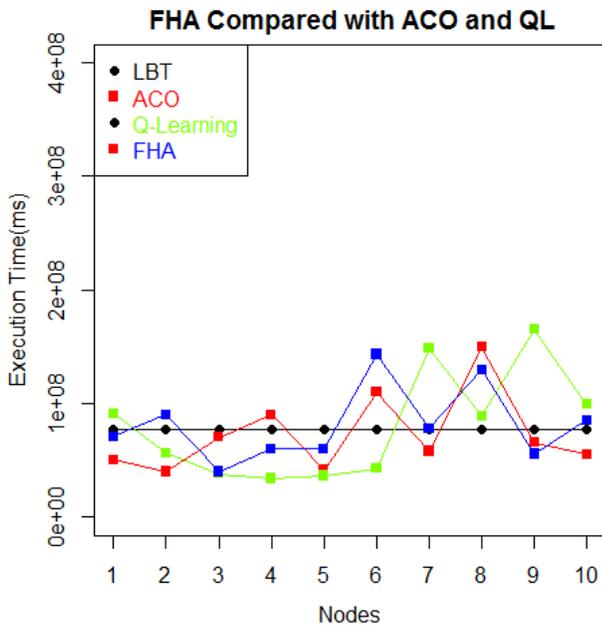


Fig. 5. Comparison Between ACO, Q-learning and FHA Curves Duration.

We use the ratio λ to assess the performance of proposed method FHA. The λ is expressed as follows.

$$\lambda_{FHA} = \frac{\theta_{FHA}}{\theta_{LBT}} \tag{17}$$

In this section, as shown in Fig. 5, the performance of FHA algorithm is tested in comparison with ACO and QL algorithms, which are two of the well-known scheduling methods for Heterogenous distributed systems as mentioned in [25, 26].

Table VII shows the ratios calculated at the end of each scheduling of 1000 tasks respectively by FHA, ACO, QL.

Clearly, it is shown that the proposed FHA method allows the dispatcher to schedule tasks among the resources much more efficiently than the ACO and QL methods.

The relevant question is whether the system will react in the same way and the same optimality to more complex and more heterogeneous tasks. The answer is of course no, we will not always find the same ratios since it is an np hard problem. Hence the need to develop the practice of artificial intelligence to avoid the repetition of the same scheduling errors and to master as possible the optimal control of all available resources. Looking forward, our goal is to be focused on deep learning through the experience accumulated during scheduling already realized, which allows the dispatcher to review continuously its strategy and then to reduce the total tardiness.

TABLE VII. CALCULATED RATIOS: λ_{FHA} , λ_{ACO} , λ_{QL}

λ_{FHA}	λ_{ACO}	λ_{QL}
1.85	1.95	2.14

- [16] Y.Hajoui, M. Youssfi, O. Bouattane and E.Illoussamen "NEW MODEL OF FRAMEWORK FOR TASK SCHEDULING BASED ON MOBILE AGENTS,". Journal of Theoretical & Applied Information Technology . Vol. 81 Issue 1, p65-72 ; October,2015.
- [17] M. Youssfi and O. Bouattane ,"Efficient Load Balancing Algorithm for Distributed Systems Using Mobile Agents ," Advanced Studies in Theoretical Physics Vol. 9, 2015, no. 5, pp.245 - 253.
- [18] A. Revar, M. Andhariya, D. Sutariya, "Load Balancing in Grid Environment using Machine Learning - Innovative Approach, " International Journal of Computer Applications (0975 – 8887), Volume 8– No.10, October 2010
- [19] TarekHelmy ,Hamdi Al-Jamimi, Bashar Ahmed, HamzahLoqman .'Fuzzy Logic–Based Scheme for Load Balancing in Grid Services', A Journal of Software Engineering and Applications, pp. 149-156. (December 2012)
- [20] A. Galstyan, K. Czajkowski, K. Lerman, Resource allocation in the grid with learning agents, Journal of Grid Computing 3 (2005) 91–100.
- [21] Coffman Jr EG, Garey MR, Johnson DS. Approximation algorithms for bin packing: a survey. Approximation algorithms for NP-hard problems, PWS Publishing Co., 1996; 46–93.
- [22] Ludwig, S.A., Moallem, A. : Swarm Intelligence Approaches for Grid Load Balancing.J Grid Computing 9, 279–301 (2011)
- [23] Kwang, M.S., Sun, H.W.: Ant colony optimization for routing and load-balancing: survey and new directions.IEEE Trans. Syst. Man Cybern. Part A33(5), 560–572(2003)
- [24] Buyya,R.,Murshed,M. :GridSim :a toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing.Journal of concurrency and computation practice and experience 14(13-15),1175-1220(2002)
- [25] Y. Hajoui, O. Bouattane, M. Youssfi, and E. Illoussamen, "New load balancing Framework based on mobile AGENT and ant-colony optimization technique". In: Proceedings of International Conference on Intelligent Systems and Computer Vision (ISCV), IEEE, Fez-Morocco (2017).
- [26] Y. Hajoui, O. Bouattane, M. Youssfi, and E. Illoussamen, "Q-Learning applied to the problem of scheduling on heterogeneous architectures". International Journal of Computer Science and Network Security, vol. 18, no. 2, pp. 153–159, 2018.
- [27] R. V. Lopes and D. Menasce, "A Taxonomy of Job Scheduling on Distributed Computing Systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 12, pp. 3412–3428, 2016.
- [28] M. I. Daoud and N. Kharma, "A high performance algorithm for static task scheduling in heterogeneous distributed computing systems,"*J. Parallel Distrib. Comput.*, vol. 68, no. 4, pp. 399–409,2008
- [29] W. E. Walsh, G. Tesauro, J. O. Kephart, and R. Das, "Utility functions in autonomic systems," inProc. Int. Conf. Autonomic Comput.,May. 2004, pp. 70–77.
- [30] J. Koodziej and S. U. Khan, "Multi-level hierarchic genetic-based scheduling of independent jobs in dynamic heterogeneous grid environment,"*Inform. Sci.*, vol. 214, pp. 1–19, 2012.
- [31] J. Ullman, "NP-complete scheduling problems," *J. Comput. Syst. Sci.*, vol. 10, no. 3, pp. 384–393, 1975.
- [32] M. Drozdowski, Scheduling for Parallel Processing, 1st ed. New York, NY, USA: Springer, 2009.

Performance Comparison of QEC Network based JAVA Application and Web based PHP Application

Sanauallah Memon¹

Software Engineering Department
Mehran UET, Jamshoro, Sindh, Pakistan

Rasool Bux Palh²

SibiSoft Technologies
Karachi, Sindh, Pakistan

Muniba Memon³

Computing Department
Indus University Karachi, Sindh, Pakistan

Hina Siddique Memon⁴

Department of Computer Science
Shah Abdul Latif University Khairpur, Pakistan

Abstract—Every organization wants to automate the manual system for moving and storing their data in particular format. A QEC department takes feedback of teacher evaluation manually from the students in the university that is somehow more difficult to maintain the record of a teacher, more cost-effective and fewer chances to generate an accurate and optimized report. The computerized system has been developed that generates an accurate and optimized report, easy to maintain the record of the teacher. Lots of possibilities are available to design and develop the application using different programming languages. We have developed a network-based JAVA application and web-based PHP application to automate the manual system of teacher evaluation. The GUI of the application contains 18 questions as per policy of HEC which will be answered by the students. After submitting the answers to questions to the server, an excel report will be ready to generate. Our primary focus is to measure the performance of the server of a network-based JAVA application and web-based PHP application. Both forms contain the same scenario, but here we have to find which form is more suitable and beneficent for an organization in terms of their server's performance parameters like average response time, throughput, and standard deviation and data transfer rate.

Keywords—QEC; network based JAVA; web based PHP; server; apache JMeter

I. INTRODUCTION

As the growing of software technology day by day, the software applications are needed for every organization to run their system correctly. Data is the bits of information, cross-region and comprehensive on the evaluation of environmental impact [1]. Framed collection of data is called database. Database management system facilitates the user to collaborate with databases for retrieving, managing and accessing the data. The database supply effective methods to stock the facts. There is research required to collar database system and to search approaches for system's knowledge-based technology [2]. By using the web interface and availability of resources and rapid growth of internet, a user can easily access the database by performing operations on it. The web has a medium containing huge data where the user can access it through the web [3]. The Internet has become an important key resource of an organization. The usage of internet service is now a matter of learning and analyzed as a dynamic process [4].

II. BACKGROUND

Data and Information are the synonym term to each other, but each contains accurate and exact meaning. There are two concepts to collect and manage the data in the database either manually or systematically. To receive and calculate the data manually and to maintain and generate the optimized report is more quite tricky.

In order to create and manage data of the QEC in an organization, there is a need for the computerized system. The automated system is an excellent solution to manage information. The aim to develop a QEC application is to automate the manual system of teacher evaluation. Many resources are required to get teachers evaluation, and lots of analysis is necessary to generate the excel report, and there were thousands of papers were distributed among students to get their teachers evaluation which was somehow more difficult to arrange a number of sets needed for assessment of the students, but this computerized system replaces the manual systems.

The two most popular programming languages JAVA and PHP contain the open source for designing and developing different applications. Comparing both languages, this paper analyzes that the interoperability support of JAVA is greater than PHP. JAVA takes more time to program, but it is a stable application, covers much security problems and gives better influences [5]. Strongly typed language such as JAVA expressed their ability to produce robust, easily maintainable applications while lightweight language such as PHP is critical to provide infrastructure for component-based applications [6].

The GUI (Graphical User Interface) of the application contains only 18 question as per policy of HEC which will be answered by the student. After submitting these answers to the server, an excel report will be ready to generate.

III. LIFE CYCLE OF QUALITY ENHANCEMENT CELL APPLICATION

To describe the possible environmental effect of the system, a life cycle assessment is performed [7]. The application mainly consists of three parts described below:

- **Server Program:** The server program is a server-side program which machine must be running. Whenever any of batches is to be evaluated for assessment of teacher and on same machine database file exists, all results submitted by the student from the client will be stored in database file using server program.
- **Client Program:** The client application is part of QEC application which will be running on LAN (local area network), i.e. (Computer Lab) to evaluate teacher using QEC application.
- **Admin Panel:** Admin panel will reside on the server side to manage the admin related task like inserting teacher, deleting teacher from the database and setting up that for which batch we are getting the evaluation from the students. Due to the significant distance between a client-side application and server-side application factor, the quality of service of network-based application may demean [8].

IV. FEATURES OF APPLICATION

- **Accuracy:** The application generates excel report which is 100% accurate as per the result submitted by the students the manual report may not be accurate because that is calculated by any human and human can make an error, but there is no any chance for error in the computer-generated report.
- **Financial Benefit:** In manual to get an assessment of any teacher, there is need of a lot of papers which were to be distributed among the students, and they provide their answer on paper after getting these all paper these were submitted to QEC, and then QEC use to start the process to generate a report.
- **Time-saving in terms of Generate Assessment Report:** The report is exported to excel file which exposes the result of all teachers of one department this report is generated within 3 seconds whereas in the manual system it may take about a number of days for a single department to generate the same report.
- **Dynamic Report:** The program written for generating the report is smart enough to create dynamic report. Dynamic means whatever result is submitted by the students the program will read a database file and prepares the report accordingly and exports that report to excel. The report is totally depending on the result being submitted by the students.
- **User-friendly Application:** Application is straightforward to use because it does not require any input from students except the answer of question all other information is maintained by application itself on runtime suppose student from any department has finished his evaluation for one teacher then he need not select another teacher only he presses Submit & Proceed button then application submits these answers to server and selects another teacher for the same students and in last when the students completed evaluation for all teachers Submit & Proceed button goes disable, and message is being displayed the students "Your Feedback submitted successful Thanks for your usual co-operation" after this that students feedback entirely submitted and save to server machine.
- **Maintain log file at server side:** Any computer connected to server program is being registered with its computer name in log file name ComputerList.log also it displays the state of the client whether the connected client has completed the teacher evaluation or still working on Performa suppose the connected client name is John then its entry in the log file will be like 01. John NO here John is the name of connected client NO means this client is still working on Performa as soon as this client completes the Performa the status of that client is changed from NO to YES. YES means this client has completed filling Performa.
- **Application is configurable:** Service-oriented architecture is developed for configurable application packages for the use [9]. Most of the parameters are configurable in demand; admin does not need to change the code at any point a file named config.properties provide the feasibility to change any parameter by changing the code base.
- **Robustness of application:** The change effects are related to the robustness which is before and after a change has occurred, scaled differences in system parameters [10]. The robustness of application maintains the client state on the server side.

A leading web based language development language PHP organizes without particular course of action about things or structure [11]. Attachment modifying is used to correspond those you quit offering on that one PC to an alternate machine with one another clinched alongside An p2p framework [12]. The java system built modifying will be very basic Furthermore not difficult Yet remains puzzling [13].

There are two phases in QEC application as shown in Table I.

- Server Application Files
- Client Application Files

These two phases are strictly worked with each other.

Database details used in the application are shown in Table II.

TABLE II. QEC APPLICATION TECHNICAL DETAILS

SERVER APPLICATION FILES				
S#	FILE NAME	FILE TYPE	LOC	PURPOSE
1.	QEC Server	JAVA	316	Start server and handles all requests coming from multiple clients.
2.	Generate Teacher Evaluation Report	JAVA	508	Generate dynamic report and export to excel
3.	Answers DTO	JAVA	87	Populate answers from database
4.	Departments DTO	JAVA	33	
5.	Teachers DTO	JAVA	44	Populate departments for database
6.	Application Utility	JAVA	40	Its utility class which reads the parameters from config.properties and provide these parameters to server program
7.	Database Manager	JAVA	303	Manages all database transaction
8.	Config.properties	Properties	2	Maintain configurable parameters of application
9.	Computer List	LOG	N/A	Logs the computer name of client connected to server
CLIENT APPLICATION FILES				
1.	Main	JAVA	23	Starts Client application
2.	Teacher Evaluation Form	JAVA	2909	User interface of client program from where student fill the teacher evaluation proforma-10
3.	Today	JAVA	109	Returns data time day and year information

V. FEATURE COMPARISON OF QEC APPLICATION

To compare the features of QEC application with the application which is already used in Shaheed Benazir Bhutto University Nawabshah, we have accepted the path of creating a list of required features.

Technical requirements of applications are shown in Table III, analyzes the applications on the basis of particular features and their availability, each feature can contain one of the following values that are: ✓ (yes, present), ✗ (no, not present), + (present but limited use).

TABLE III. DATABASE DETAILS

S#	TABLE NAME	No: of Rows	No: of Columns
1.	Department	3	2
2.	Teachers	9	5
3.	Answers	288	7
4.	MAINTAIN-EACH-CLIENT-STATE	23	4

VI. ADVANTAGES OF DESKTOP AND WEB-BASED APPLICATIONS

The advantages of desktop based and web-based application are shown in Table IV.

TABLE IV. FEATURE COMPARISON OF QEC APPLICATION

CATEGORIES	FEATURES	QEC APP	QEC APP OF SBBU
Interface	Login history	✓	✗
Performance	Page Load	✓	✓
Dynamic Report	Excel report	✓	✗
Accuracy	Error Calculation	✗	+
Development	Deployment	✓	+
Maintenance	Log File	✓	✗

TABLE V. ADVANTAGES OF DESKTOP AND WEB-BASED APPLICATION

S.No	DESKTOP BASED APPLICATION	WEB BASED APPLICATION
(1)	No reliance on an internet	Cross Platform
(2)	Much easier to customize	Huge Community
(3)	High Efficiency	Fast Development Cycle
(4)	User Interface Flexibility	Standard Based

TABLE VI. DISADVANTAGES OF DESKTOP AND WEB-BASED APPLICATION

S.No	DESKTOP BASED APPLICATION	WEB BASED APPLICATION
(1)	Speed at which Software upgrades	Larger Overhead
(2)	Restricted to a single standalone machine	Less control over computer resources
(3)	Less Connectivity	Accessible any where

The disadvantages of desktop based and web-based application are shown in Table V.

VII. DATABASE SCHEMA

In the framework of the traditional information system, database developing is at the center of an information system expresses a hard alimony problem [14]. Database Schema expresses the logical view of the whole database that describe show to manage and organize the data or information, and the relationship between them. It also explains all the set of rules that are to be enforced on data. A database schema defines its entities, attributes and the relationship between them. It contains descriptive details of the database that can be illustrated by the schema. A database designer designs a database schema that helps the developer or programmer to manage and understand the database and to make a useful database.

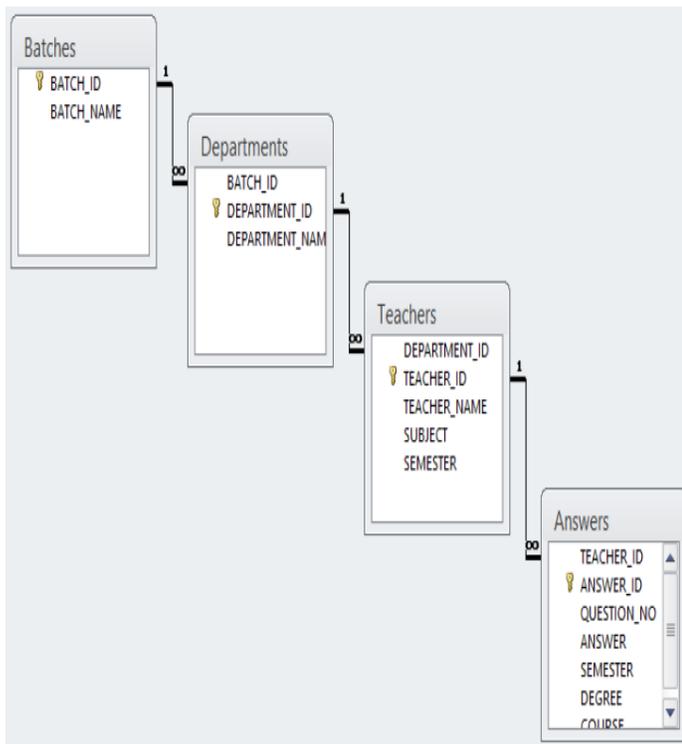


Fig. 1. Database Schema of the Application.

The databases used in QEC application have four tables, i.e., batches, departments, teachers and answers. These four tables have one to many relations with each other as shown in Fig. 1.

VIII. QEC PROJECT DESIGN

Designing a project plays an important role which follows the methods and theories that contest the behavior of end users and project goals [15]. The project is designed and developed to provide the advanced functionalities of feedback of teacher evaluation, which contains some teachers related questions as per policy by HEC and answered by the students during the assessment of feedback of teacher evaluation. This feedback evaluation system generates an accurate and optimized report in excel and provides correct results.

IX. QEC APPLICATION MAIN SCREEN

This is the main view of QEC application that shows text boxes, check boxes, buttons and different contents of the application as shown in Fig. 2.

The text boxes contain specific information of the department's related Instructors and course titles which is handled by the system administrator. These boxes are already selected by the administrator when taking feedback of teacher evaluation from the students of the different batches of the departments. There is no any facility to make any changes in the text boxes to the student. The checkboxes show different options for each question. Each question contains some teacher's related meaning. The Graphical User Interface of this application provides only 18 items as per policy of HEC related with the teacher which will be answered by the student. By attempting all the questions of the teacher evaluation form, press Submit and proceed. The instructor name and course title have been changed dynamically after attempting and submitting all the questions related to the teacher and his course title. There is validation in attempting the entire questions. The submitting will not proceed till trying all the questions. The grading scale of each Instructor has been generated in excel file after submitting and proceeding the application. The request from QECClient.JAVA sends to the server, the server loads that file obtained from the client, the file loads all the system's data i.e., Personal computer name for security purpose on the server and interact with the database. When all the data load on the server successfully, the server generates the response and sends to the client in the shape of Message Dialog Box. Fig. 3 shows the flow of the main screen.

When the student fills the teacher evaluation form completely, the entire teacher's related data loads on the server. Then an excel report is generated on the server which contains the grading scale of every teacher of the specific department which is already being selected by the administrator. The excel report makes accurate and optimized information of the teacher.

Fig. 2. QEC Main Screen.

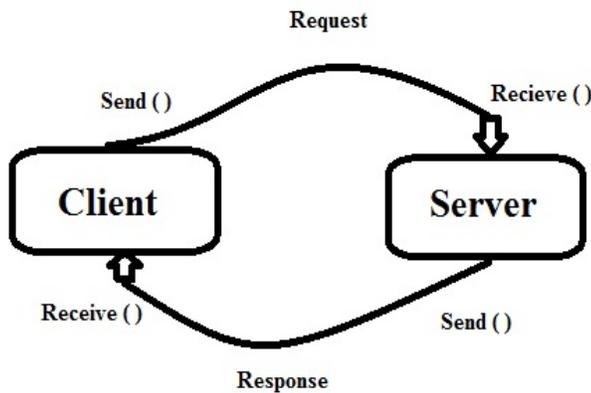


Fig. 3. Flow of the Main Screen.

X. ADMIN PANEL SCREEN

Admin panel plays a vital role in updating and managing relevant data or information of the teacher evaluation feedback through web administrator. The administrator uses admin panel screen which contains admin login module. The module provides two textboxes, his/her login id, and password as shown in Fig. 4.

When the administrator successfully login, the control panel screen is visible on the mainscreen which yields the provision to manage QEC application. It produces the facility to enter the facts and statistics collected together for reference of particular batch whose feedback is to be taken. The admin makes changes in department name, batch name, and the name of instructor and course titles. It provides an interface to the

client where all these mentioned above parameters are pre-selected. There is no any chance for the client to makes any changes in the above parameters.

Fig. 5 shows the movement of client’s data which shows how data flows from dashboard screen to the server. The data is correctly filled by the client and sends a request to the server. The server stores all the data of the client into the database and generates a response. There is a valid option when filling up the feedback teacher evaluation form. When the data is successfully inserted into the database then it creates a message to the client like this “The Data is Successfully Saved.”



Fig. 4. Circulate of the Panel Display.

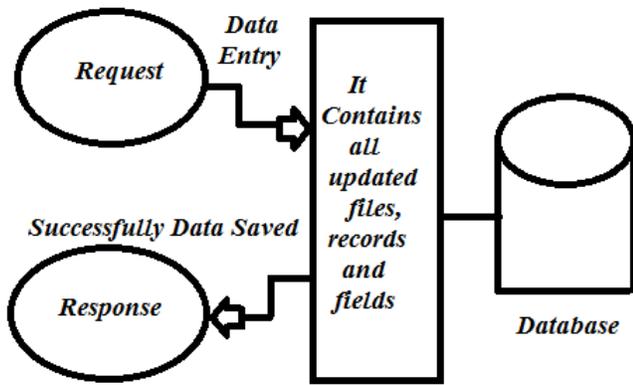


Fig. 5. Flow of Data from Control Panel / Dashboard to Database.

XI. EXPERIMENTS AND RESULTS

There is need to design and develop QEC applications for taking the proper feedback of teacher evaluation of the departments. The applications are developed in JAVA and PHP computer programming languages. There is too much obligatory to enable the run of applications. Keep application remains in a proper state; the architecture is designed properly of both the applications. There are certain efforts have been needed for obtaining the final results of performance parameters such that design and develop the application. Installation of both applications, JAVA desktop based application, is installed on each machine and web components are placed at the server. To connect each machine with the server, the paper contains same URL Address [IP address / index.PHP]. The paper enables to fetch same size of organization (here same number of students) that may be offline (not connected with internet) locally. To measure the performance parameters of both the applications in terms of average response time, throughput, standard deviation and data transfer rate, there is need of a load balancing testing tool. There are lots of testing tools available for measuring the performance of the desktop application and web-based application. The Apache JMeter testing tool is one among all of them. Jmeter is a popular and powerful free load testing tool which is used for measuring the performance parameters of the applications [16]. The Apache JMeter is the open source testing tool which is used to measure the performance of the server by using the applications. It acts by a surrogate as client side of client/server applications. The server resources like CPU loads, memory loads and response time, is calculated by Jmeter [17]. There is a need to connect load balancing testing tool (Apache Jmeter) with the Server to measure the different performance parameters.

XII. HOW JMETER WORKS

Jmeter contains a graphical user interface with the clear test plan. To conduct a test, there is a need of test plan in JMeter which describes the steps will take to perform some specific testing task [18]. A test plan is made up of a sequence of test components which determines how the load will be simulated. First, there is a need to build a basic test plan as shown in Fig. 6.

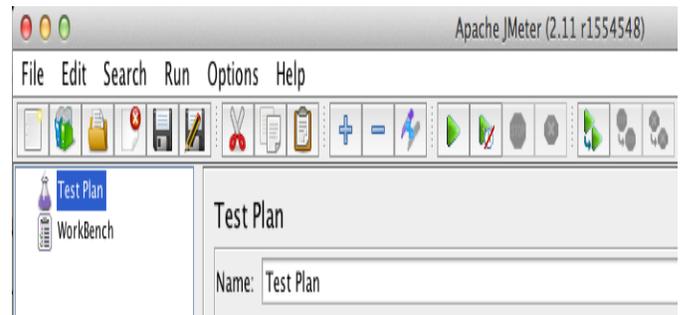


Fig. 6. Build a Basic Test Plan.

The JMeter tool is used to measure the variety of tests which are useful for the environment. A test plan includes the thread group element which is used to identify clearly and definitely the number of running threads and ramp up period [19]. A thread group contains some useful properties that influence the load test such as threads (users), ramp up period and loop count. The loop count shows the number of times, the test will be executed [20]. Here 1 loop count means each test will be repeated 1 time. The threads contain the number of users which you want to simulate, here we set 50 users. Ramp up period is the duration of time will divide the start of thread over, here we set 10 seconds. Loop count shows the number of times, the loop will be executed; here we set 1 count as shown in Fig. 7.

In test plan, The HTTP default config element is used to set the HTTP default request's values. This parameter is more useful to measure multiple HTTP requests on the same server. If the web server runs on the local machine, then there is no need to set the IP address as shown in Fig. 8.

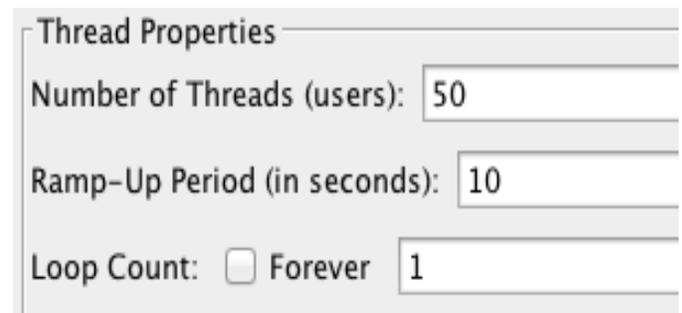


Fig. 7. Thread Properties.

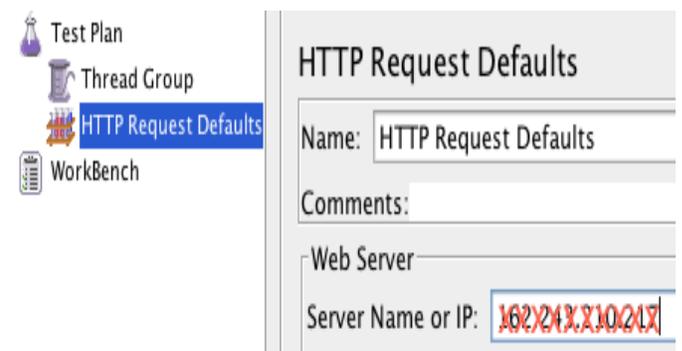


Fig. 8. Http Request Default.

XIII. SUMMARY REPORT IN APACHE JMETER

The summary report in apache jmeter contains a list of performance parameters which we get after performing some series of steps in apache jmeter. The summary report shows the overall response of the server after sending the requests. The performance parameters are listed below:

- 1) *Label*: The label section displays all the recorded HTTP requests.
- 2) *Samples*: Samples section shows the number of HTTP requests or the number of users/ threads during the test run.
- 3) *Average response time*: Average response time refers to the amount of time; the server takes and returns the response to the user. It is the average response time of HTTP requests and is calculated in milliseconds.
- 4) *Min average response time*: It is a minimum amount of response time obtained from the server after sending the HTTP requests. It is also calculated in milliseconds.
- 5) *Max average response time*: It is the maximum amount of response time received from the server after sending the HTTP requests. It is too calculated in milliseconds.
- 6) *Standard deviation*: It is a quantity which shows how much group members are to be dissimilar from the signify value. A low standard deviation value is also called expected value which expresses the data points are close to the mean value.
- 7) *Error rate*: In samples during the run, it shows the error percentage.
- 8) *Throughput*: It defines the number of requests per unit of time that the claims which are sent from clients to the server during the test and is calculated either in seconds, minutes or hours.
- 9) *Data transfer rate*: It shows the amount of speed through which the data can be emitted from one device to another device and is calculated in megabits or megabytes. Throughput is also another word used for Data Transfer rate.

XIV. PERFORMANCE EVALUATION

Performance evaluation is a key task to expect when designing and implementing several technologies. Few common complications might occur due to implementation of either lightweight or large IT environments. The load balancing testing tools can be useful computing several parameters of performance of the applications.

A. Performance Evaluation of PHP Application

The following are the results of QEC application in apache jmeter, which is being designed and developed in PHP programming language. As the application running on the single machine, give the localhost address and port number (8080) when using the apache jmeter. This research paper explains that 300 numbers of samples/users are sent to the server and get the response from the server in the following performance parameters results in apache jmeter.

- Number of Samples/Requests = 300 requests
- Average Response Time = 33 ms = 0.033sec

- Min Response Time = 4 ms = 0.004sec
- Max Response Time = 384 ms = 0.384sec
- Standard Deviation = 91.71
- Error rate = 0.00%
- Throughput = (number of requests) / (total time) = 31.3 / sec
- Data Transfer Rate = KB/sec = 34.81

B. Performance Evaluation of JAVA Application

The application which is designed and developed in JAVA programming language shows the response in different parameter results of the server after sending the request of JAVA QEC application in apache jmeter. Socket programming is used inside JAVA because the JAVA application is network based. As the application running on the single machine, an object from the localhost address was sent by using a port number (9090), 300 number of samples/users to the server and get the response back from the server in the following performance parameters results in apache jmeter.

- Number of Samples/Requests = 300 requests
- Average Response Time = 22 ms = 0.022sec
- Min Response Time = 4 ms = 0.004sec
- Max Response Time = 367ms = 0.367sec
- Standard Deviation = 61.83, Error rate = 0.00%
- Throughput = (number of requests) / (total time) = 31.2 / sec
- Data Transfer Rate = KB/Sec = 34.64

XV. COMPARING OF PHP AND JAVA APPLICATION IN TERMS OF PERFORMANCE PARAMETERS

To compare the graphs, and the best practices help the designers and programmers for selecting the visualization techniques and also allows users to interpret and perform related tasks of how graphs are used to gather. The graph which select must support the user's primary and secondary tasks. If the overall data support these tasks with the same measurement unit than plot both number of events in a single graph. The data or information which changes continuously over time, the Line Graph is a better approach to increase or decrease in data and to display the data or information. A commonly used, line graph contains a series of data points which are connected through the straight line on two axes and changes over time.

- Line Graph in Terms of Average Response Time

Here the paper compares the server response in line graphs of both PHP and JAVA applications in terms of Average response time. Average response time shows the average amount of time that the server takes, a client must wait before request before getting a response from the server. Different response of server had gotten after sending the requests to the server time by time. By using PHP application, sent 50 number of requests or samples to the server, got average response time in milliseconds which is 48.75 ms. This paper defines when

sent 300 number of requests to the server, got another result which is 33 ms. In the same way, by using the network-based JAVA application when sent 50 number of requests to the server, got 43 ms of average response time, but when sent 300 numbers of requests to the server, got different average response time which is 22 ms as shown in Fig. 9.

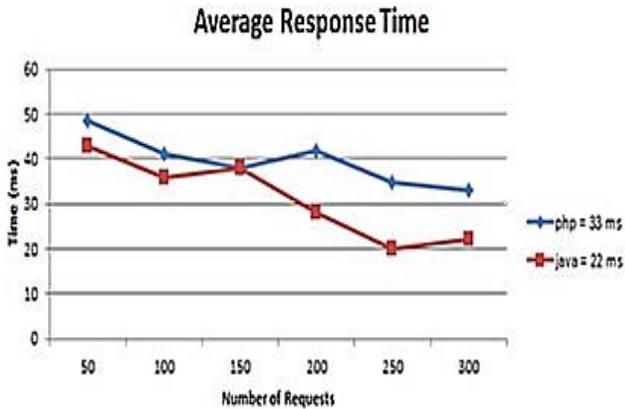


Fig. 9. Line Graph in Terms of Average Response Time.

- Line Graph in Terms of Throughput

Throughput is defined as the measure of units of information; the system can process in a specific period of time or in other words, it can be described as the number of requests per unit time. Here it is calculated in seconds. The different throughput of the server had gotten after sending the requests to the server time by time. By using PHP application, when sent 50 numbers of requests or samples to the server, got throughput in seconds which is 5 seconds. When sent 300 numbers of requests to the server, obtained 31.3 seconds. In the same way, by using the network-based JAVA application when sent 50 numbers of requests to the server, get 5 seconds of throughput, but when sent 300 numbers of requests to the server, got different throughput which is 31.2 seconds. There is little bit variation in performances of both applications, as shown in Fig. 10.

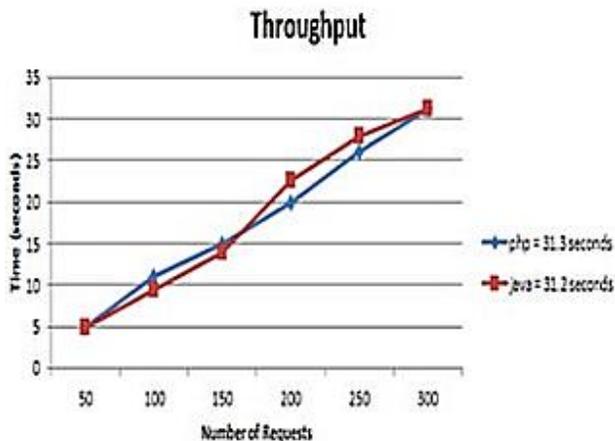


Fig. 10. Line Graph in Terms of Throughput.

- Line Graph in Terms of Standard Deviation

This paper compares the server response in the line graph of both PHP and JAVA applications in terms of standard deviation. Standard deviation is defined as a measure which is used to liberate the amount of difference of a lay of data values. The data points close to the expected value express the low standard deviation. Different standard deviation values of the server have gotten after sending the number of requests to the server time by time. By using PHP application, when sent 50 numbers of requests or samples to the server, got standard deviation which is 100 data values. When sent 300 numbers of requests to the server, obtained 91.71 data values. In the same way, by using the network-based JAVA application when sent 50 numbers of requests to the server, got 78 data values of standard deviation, but when sent 300 numbers of requests to the server, got different standard deviation which is 61.83 data values, as shown in Fig. 11.

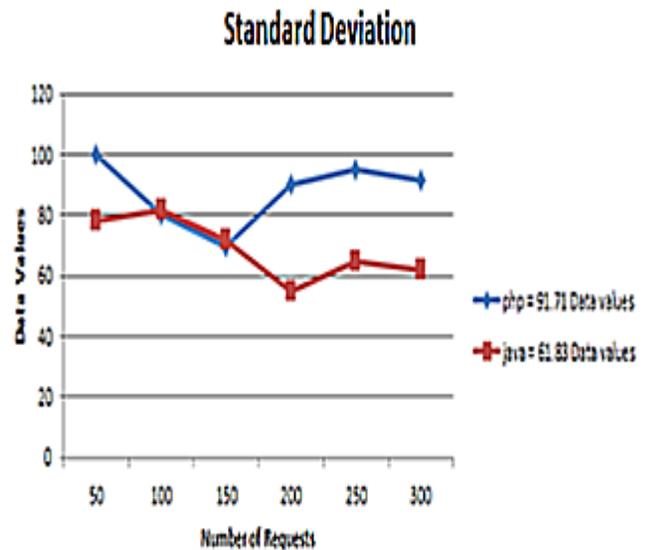


Fig. 11. Line Graph in Terms of Standard Deviation.

- Line Graph in Terms of Data Transfer Rate

This paper compares the server response in the line graph of both PHP and JAVA applications in terms of data transfer rate. The data transfer rate shows the speed of transmitting the data from one device to another device. It is measured in megabits or megabytes per second. Different data transfer rate values from the server have gotten after sending the number of requests to the server. By using PHP application, when sent 50 numbers of requests or samples to the server, got 40 kb per second of data transfer rate. When sent 300 numbers of requests to the server, obtained 34.81 kb per second of data transfer rate. In the same way, by using the network-based JAVA application when sent 50 numbers of requests to the server, get 48 kilobytes per second of data transfer rate, but when sent 300 numbers of requests to the server, got different data transfer rate which is 34.64 kb per second, as shown in Fig. 12.

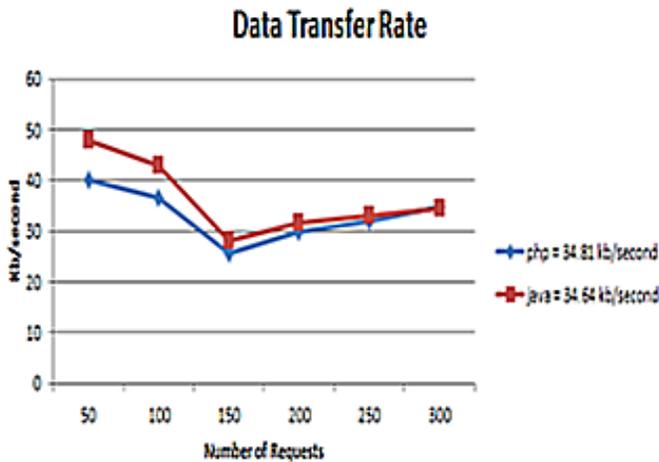


Fig. 12. Line Graph in Terms of Data Transfer Rate.

XVI. ACHIEVEMENTS

All the objectives achieved successfully which are:

- Design and Develop QEC Network Based JAVA Application
- Design and Develop QEC Web Based PHP Application
- Compare the results of both Applications in terms of Average Response Time, Throughput, Standard Deviation and Data Transfer Rate Parameters of the server.

In the results of objectives, obtained the desired aim that is:

- To Find the most feasible solution for getting feedback data for QEC

XVII. CONCLUSION

Now a day, the software technology is growing rapidly day by day. Each and every organization provides facility to the user to run their system correctly takes less time and gives better and accurate results which are beneficent for the organization. All the required information is collected from the clients and has to know which type of application is needed for the organization. The software architects provide proper solution according to the user requirements. By comparing the server results of network-based JAVA application and web-based PHP application in terms of performance parameters, i.e., average response time, throughput, standard deviation and data transfer rate, it is observed that JAVA application is the best feasible solution for feedback evaluation of the organization. With the results of average response time and throughput, we concluded that the JAVA application gives quick response than PHP.

The network-based JAVA application contains fewer data values of the standard deviation as compared to the web-based PHP application. By using the apache jmeter and with the core study of standard deviation, concluded is that the application has fewer data values of standard deviation will give the better influences. So JAVA application is also better than PHP in terms of standard deviation, it contains less standard deviation. The apache jmeter also keeps the feature of data

transfer rate and calculate it in kilobytes per seconds. There is a little bit different but almost same in data transfer rate of both the applications. By comparing the results of each and every parameter, it is concluded that network-based JAVA application is better than web-based PHP application in all aspects for the organization.

In future, the educator's testament assessment will make more by utilizing these sorts about electronic framework requisitions. It holds ton of characteristics and gives great deal of profits of association such provides for fast reaction what's more generates exact what's more optimized report card about educator's testament. This electronic framework gives preferred comes about what's more certain ways appears to be will be more positive position over manual framework for QEC.

XVIII. DISCUSSION

It may be as a relatable point to each and each association moves towards on the most recent innovations whose use will be ended up additional beneficent for future worth of efforts that is whichever oversaw economy system, majority of the data system, sound alternately feature imparting also a great deal a great amount. Each association needs on move starting with manual framework should programmed framework by utilizing the web built innovations. Our research inspiration is related on the execution from claiming QEC requisitions uncommonly outlined also created over system built JAVA provision and web based PHP provision. The provisions hold the same eighteen inquiries viewing with those educator's testament assessment concerning illustration for every strategy for higher education commission gives the web submission about educator's testament assessment manifestation. In the event for outline judgment report, both requisitions produce exact what's more optimized report card. Anyway here fundamental objective will be should measure the server execution of both requisitions as far as execution parameters like Average response time, throughput, standard deviation and data transfer rate. The load balancing testing tool Apache JMeter is selected for getting the final performance parameters results. The number of requests is sent to the server. When the requests successfully received by the server generate response for the client in different measurement units in milliseconds, seconds and microseconds.

XIX. ACKNOWLEDGMENT

This research has been conducted in Mehran University of Engineering and Technology (MUET) as part of Master's dissertation.

REFERENCES

- [1] Liang, Zhu, et al. "ArcObjects-based eco-environmental data management information system for Three Gorges Project." IEEE International Conference on Information Technology and Computer Science (ITCS 2009), Vol. 2.
- [2] Brodie, Michael L., and John Mylopoulos, eds. On knowledge base management systems: integrating artificial intelligence and database technologies. Springer Science & Business Media, 2012.
- [3] Ghanem, Thanaa M., and Walid G. Aref. Databases deepen the web. IEEE Computer 37.1 (2004): 116-117.
- [4] Van Dijk, Jan AGM, Oscar Peters, and Wolfgang Ebbers. "Explaining the acceptance and use of government Internet services: A multivariate

- analysis of 2006 survey data in the Netherlands." *Government Information Quarterly* 25.3 (2008): 379-399.
- [5] Walden, James, et al. "Idea: JAVA vs. PHP: security implications of language choice for web applications." *International Symposium on Engineering Secure Software and Systems*. Springer Berlin Heidelberg, 2010.
- [6] Wright, William, and Dana Moore. "Agile language development: the next generation." *IEEE Aerospace Conference*, 2006.
- [7] Moberg, Åsa, et al. "Printed and tablet e-paper newspaper from an environmental perspective, A screening life cycle assessment." *Environmental Impact Assessment Review* 30.3 (2010): 177-191.
- [8] Hoshino, Yuta, et al. "An on-line algorithm to determine the location of the server in a server migration service." *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, 2015.
- [9] Mietzner, Ralph, Frank Leymann, and Mike P. Papazoglou. "Defining composite configurableSaaS application packages using SCA, variability descriptors and multi-tenancy patterns. Third IEEE International Conference on "Internet and Web Applications and Services, 2008 (ICIW'08).
- [10] Ross, Adam M., Donna H. Rhodes, and Daniel E. Hastings. "Defining changeability: Reconciling flexibility, adaptability, scalability, modifiability, and robustness for maintaining system lifecycle value." *Systems Engineering* 11.3 (2008): 246-262.
- [11] Cui, Wei, et al. "The research of PHP development a framework based on MVC pattern. Fourth IEEE International Conference on Computer Sciences and Convergence Information Technology, ICCIT'09.
- [12] Loo, Alfred Wai-Sing. "JAVA Network Programming." *Peer-to-Peer Computing*. Springer London, 2007, 67-90.
- [13] Harold, Elliotte Rusty. *JAVA network programming*. "O'Reilly Media, Inc.", 2013.
- [14] Curino, Carlo A., et al. "Schema evolution in Wikipedia: toward a web information system benchmark." In *International Conference on Enterprise Information Systems (ICEIS)* 2008.
- [15] Pernencar, Cláudia, Teresa Romão, and Graça Rocha Simões. "The design process of an e-Health project: Applying the HSI framework for interface analysis." *Serious Games and Applications for Health (SeGAH), 2016 IEEE International Conference on*. IEEE, 2016.
- [16] JMeter, Apache. "Apache software foundation." (2010).
- [17] Team, JMeter. *Enhancement of JMeter*. Diss. Indian Institute of Technology, Bombay Mumbai, 2013.
- [18] Kurniawan, Budi, and JMeteris a JAVA. "Using JMeter." <http://onJAVA.com/Ipt/a/3066>, (Jan. 15, 2003) (2003):1-9.
- [19] Nevedrov, Dmitri. "Using JMeter to Performance Test Web Services." *Published on 57 dev2dev* (<http://dev2dev.bea.com/>) (2006).
- [20] Alsmadi, Izzat, and SaschaAlda. "Simulation Based Load Testing In Web Services." *The 6th International Conference on Information Technology* May, 2013. Vol. 8.

Aspect-Combining Functions for Modular MapReduce Solutions

Cristian Vidal Silva¹, Rodolfo Villarroel², José Rubio³, Franklin Johnson⁴, Érika Madariaga⁵, Alberto Urzúa⁶, Luis Carter⁷, Camilo Campos-Valdés*⁸, Xaviera A. López-Cortés⁹

¹Ingeniería Civil Informática, Escuela de Ingeniería,
Universidad Viña del Mar, Viña del Mar, Chile

²Escuela de Ingeniería Informática, Facultad de Ing.,
Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

³Área Académica de Informática y, Telecomunicaciones,
Universidad Tecnológica de Chile, INACAP, Santiago, Chile

⁴Depto. Disciplinario de Computación e Informática,
Facultad de Ingeniería, Universidad de Playa Ancha, Valparaíso, Chile

⁵Ingeniería Informática, Facultad de Ingeniería,
Ciencia y Tecnología, Universidad Bernardo O'Higgins, Santiago, Chile

⁶Escuela de Kinesiología, Facultad de Salud,
Universidad Santo Tomás, Talca, Chile

⁷Ingeniería Civil Industrial, Facultad de Ingeniería,
Universidad Autónoma de Chile, Talca, Chile

⁸Programa Doctorado en Sistemas de Ingeniería, Facultad de Ingeniería,
Universidad de Talca, Curicó, Chile

⁹Depto. de Computación e Industrias, Facultad de Ingeniería,
Universidad Católica del Maule, Talca, Chile

Abstract—MapReduce represents a programming framework for modular Big Data computation that uses a function map to identify and target intermediate data in the mapping phase, and a function reduce to summarize the output of the map function and give a final result. Because inputs for the reduce function depend on the map function's output to decrease the communication traffic of the output of map functions to the input of reduce functions, MapReduce permits defining combining function for local aggregation in the mapping phase. MapReduce Hadoop solutions do not warrant the combining functioning application. Even though there exist proposals for warranting the combining function execution, they break the modular nature of MapReduce solutions. Because Aspect-Oriented Programming (AOP) is a programming paradigm that looks for the modular software production, this article proposes and apply Aspect-Combining function, an AOP combining function, to look for a modular MapReduce solution. The Aspect-Combining application results on MapReduce Hadoop experiments highlight computing performance and modularity improvements and a warranted execution of the combining function using an AOP framework like AspectJ as a mandatory requisite.

Keywords—Combining; Hadoop; MapReduce; AOP; AspectJ; aspects

I. INTRODUCTION

MapReduce represents a computation framework aiming to solve Big Data and Big Computation issues [1]–[4]. Hadoop is a MapReduce application tool [4], [5] with two main components, the Hadoop Distributed File System (HDFS) for an 'Infrastructural' point of view and MapReduce for the 'Programming' aspect. Hence, HDFS is a distributed

and scalable file system designed for running on clusters of commodity hardware. HDFS follows the write-once, read-many approach to store huge files using streaming data access patterns to enable high throughput data access and simplifies data coherency issues [4], [5]. HDFS abstracts developers of distribution, coordination, synchronization, faults and failures, and supervision tasks details. Thus, developers must focus on two main computation functionalities: map and reduce.

Aspect-Oriented Programming (AOP) corresponds to a programming methodology for isolating crosscutting concerns functionalities and data to look for modular solutions [6]. Ideas of obliviousness and advisable classes appear in AOP. Wampler [7] indicates and demonstrates the AOP support and refinement of Object-Oriented Design (OOD) principle such as the Single Responsibility Principle (SRP) and Open-Closed Principle (OCP) mainly to remark the AOP practical benefits.

Even though MapReduce represents a framework to isolate a programmer of traditional faults and issues on traditional distributed programming approaches and frameworks, MapReduce demands to figure out solutions using their main two functions: map and reduce. Thus, these functions can include code out of their inner nature which are clear crosscutting concerns examples according to good modular programming and AOP principles [7].

Hadoop allows the definition of the combining function on the map output [5], [8], [9] to optimize the MapReduce framework functioning for local aggregation in the map phase, that is, a function to aggregate data in the map phase before sending them to the reduce phase. Even though the combiner function

is an optimization, Hadoop does not provide a guarantee of how many times it will call defined combining functions [8]. Thus, as a guarantee of combining execution, [8] proposed the use of the 'In-Mapper' Combining function, i.e., the combining function behavior directly inside the map function. Nonetheless, this solution does not respect object-oriented modularity principles such as the SRP [7], [10]. Looking for a modular application of the MapReduce programming framework, this article proposes and exemplifies the use of Aspect-Combining, an AOP application on MapReduce for the combining functions definition. Thus, the main contributions of this article are:

- Giving a review of performance and modularity issues of MapReduce combining solutions.
- Locating and justifying the presence of crosscutting-concerns in current optimal combining solutions.
- Defining and testing Aspect-Combining functions on classic case studies for getting more modular and usually more efficient results.
- Establishing the bases for future works about the symbiosis of Big Data and AOP solutions.

This article is organized as follows: Section II gives a description of the MapReduce framework and its main components. That section also explains the primary structure and principles of traditional AOP-AspectJ solutions. Section III reviews previous 'In-Mapper' Combining function and identifies crosscutting concerns issues to define Aspect-Combining functions. Section IV defines hypothesis and variables to measure in the experiments, and presents results of the use of Combining, 'In-Mapper' Combining, and Aspect-Combining proposal on a few application examples to highlight the main practical pros and cons of the Aspect-Combining function. Section V discusses validity of the established hypothesis. Section VI concludes and presents future research work.

II. MAPREDUCE AND AOP

A. MapReduce

MapReduce is a programming model proposed by Google [1]–[3] for distributed computation on massive amounts of data (Big Data), that is, MapReduce is an execution framework for large-scale data processing on clusters of commodity servers. MapReduce has already enjoyed widespread adoption by the use of Hadoop, a open-source implementation of MapReduce [5], [8].

MapReduce can refer to three concepts: 1) a programming model; 2) an execution framework to coordinates the execution of programs written in this programming style; 3) the implementation of 1) and 2), that is, MapReduce is the implementation of a programming model and its execution framework. Google is the proprietary of MapReduce implementation [1]–[3], and Hadoop is an open-source analogue substitute [5], [8].

Hadoop applies the Hadoop Distributed File System (HDFS), a highly fault-tolerant and distributed file system able to run on commodity hardware [4], [5], [8]. HDFS provides high throughput access to application data. HDFS is suitable for applications with large data sets such as the set of valid configurations in a Software Product Line (SPL) [9].

MapReduce basic idea is to partition a large problem into smaller sub-problems possibly independent able to run in parallel by different workers, that is, either by threads in a processor core, cores in a multi-core processor, multiple processors in a multi-processors machine, or many machines in a cluster [4], [8]. Fig. 1 shows the Hadoop functioning architecture. Hence, an iteration of a Hadoop solution (a.k.a job) normally executes in four steps: 1) Slicing to split the source data in multiples splices and deliver them to each map-worker or mapper. 2) Map to process the data (each mapper processes one or more chunks of data and sends the results to the shufflers). 3) Shuffle to organize the data. 4) Reduce to compact and write back results to the disk. Thus, intermediate results from each worker are then combined to yield the final output.

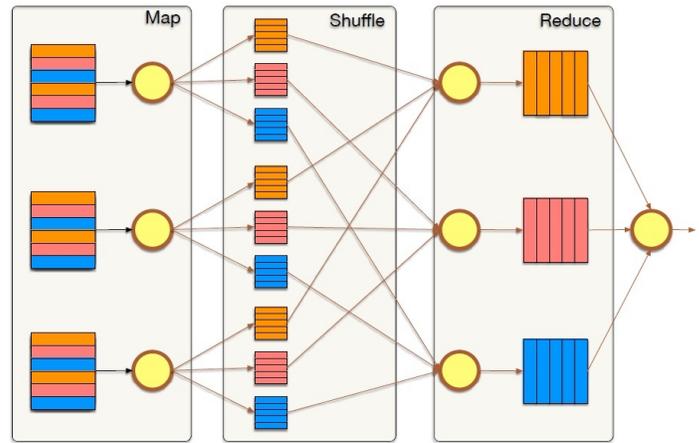


Fig. 1. MapReduce functioning architecture.

MapReduce allows for commutative and associative map functions to define combining function, that is, to decrease the amount of data shuffling between map-workers and reduce-workers [5], [8]. Combining functions work on the map functions output; hence, the output of combining functions represent the input of reduce functions. The MapReduce [1]–[3] execution framework coordinates functioning of map-workers and reduce-workers.

Hadoop solutions usually enable for the definition of a set of dependent jobs, i.e., the output of one job is used as an input for others and so on. Thus, a set of key-value records (K_{in}, V_{in}) is the input of a map function, and a $list(K_{inter}, V_{inter})$ corresponds to its output, that is, the input of a combining function if it were defined or input for the shuffling process. As was mentioned, the input for combining functions corresponds to the output of mappers, and the combining functions output will be the input for the shuffling process. Shuffling process orders and distributes data for reduce functions, that is, they get $(K_{inter}, list(V_{inter}))$ as input to produce an output (K_{out}, V_{out}) which can be the input of other map functions, and so on. Fig. 2 illustrates this described process.

B. Aspect-Oriented Programming

Aspect-Oriented Programming (AOP) [6] permits modularizing crosscutting concerns in base classes as aspects in Object-Oriented Programming (OOP). Aspects advise classes

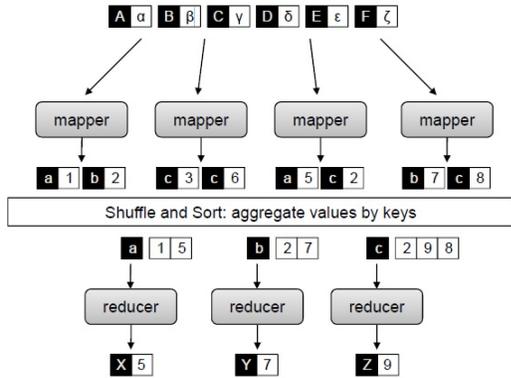


Fig. 2. A Simplified view of MapReduce.

statically in defined advisable modules and dynamically like events. AOP like AspectJ [6] defines oblivious advisable classes and modularizes crosscutting concerns as aspects, that is, orthogonal methods which are not part of the nature of advisable classes.

AOP well modularizes homogeneous crosscutting concerns as aspects [6], [7], [11]–[13]. However, aspects do not reflect the structure of refined features and the classes cohesion for the modularization of classes collaboration [14], [15]. Moreover, AOP languages like AspectJ [13], [16], [17] introduce implicit dependencies between aspects and advisable classes [18]–[21]. Hence, first, aspects do not respect the information hiding principle because oblivious classes can experience unexpected behavior and properties changes, and second, changes on the firm of advisable behavior can generate spurious and non-effective aspects. Thus, aspects need to know structure details about the advisable behavior and classes, a great issue for independent development.

Next, this article describes main AOP elements.

C. Join points and Pointcuts

A join point represents an event in the execution control flow of a program, that is, “a thing that happens” [13], [16]. Hence, in AOP [6], [11], a join point is a point of the program execution in which aspects advise advisable base modules. Examples of join points in AspectJ are method calls, method executions, object instantiations, constructor executions, field references and handler executions

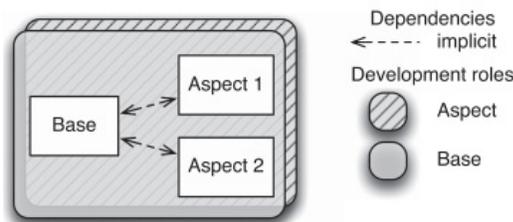


Fig. 3. AspectJ components and functioning.

```
public class HelloWorld {
    public static void main(String[] args){
        say("Hello");
        sayToPerson("Hello", "Cristian");
    }
    public static void say(String message) {
        System.out.println(message);
    }
    public static void sayToPerson(String message,
        String name) {
        System.out.println(name + ", " + message);
    }
}

public aspect BasicAspect {
    pointcut callSayMessage():
        call(public static void HelloWorld.say*(..));

    before() : callSayMessage() {
        System.out.println("Good day!");
    }
    after() : callSayMessage() {
        System.out.println("Thank you!");
    }
}
```

Fig. 4. AspectJ HelloWorld example.

According to [6], [7], a pointcut is a rule to pick out and define the join points occurrence and expose data from the execution context of those join points. Possible components of pointcut rules definition are call (method pattern), execution (method pattern), get (field pattern), set (field pattern), identifiers of time for advisable methods, objects associated to an advisable method, among others.

Just, for the pointcut definition in AOP like AspectJ languages of a method execution, two important times exist: when a methods is called (*call* time) and when a method is in execution (*execution* time). Furthermore, we can differentiate between *target* and *this* objects on the join point event, that is, the object whose method is in execution and the object that executes the method on the target object. Thus, *this* and *target* are the same object for pointcut rules of *execution* methods, and for *call* pointcut *this* is the object that order the *target* method execution.

D. Inter-type Declarations and Advices

In essence, inter-type declaration statically injects changes on fields, properties, and methods into existing advisable classes in AOP [6].

Advice defines crosscutting behavior regarding pointcut. Three type of advice in traditional AOP exist [6]: *before*, *after*, and *around* which determine how an advice runs at every picked out join point. These kinds of advice determine how the code injection works over the join points. Thus, in AOP like AspectJ languages there exist advice instances which run before their join points, run after their join points, and run in place of (or “around”) their join points.

Fig. 3 [22] details the AspectJ components and functioning structure, that is, aspects advise oblivious base modules and they present implicit dependencies among them.

Fig. 4 [11] illustrates a basic AspectJ example, an advisable class *HelloWorld* and an aspect with two advice instances to inject behavior into the advisable class before and after calling a void method that starts with the word *say* in the class *HelloWorld*.

AspectJ mainly looks for modular solutions and respecting modularity principles [16]–[18], [23]. This paper looks for getting modular MapReduce solutions by the use of AspectJ on Hadoop solutions.

III. GROUPING DATA LOCALLY IN MAPREDUCE

The MapReduce computation in Hadoop does not require to put attention on embarrassingly-parallel issues such as synchronization and deadlock [7], [8]. Hadoop and MapReduce solutions possibly involve large data-intensive transferring from map-worker to reduce worker instances. Thus, since data transferring can be of a high cost; for the *local aggregation*, combining functions can considerably diminish the map output records with the same key in the map-workers.

```
public class WordCountMapper extends MapReduceBase
implements
Mapper<LongWritable, Text, Text, IntWritable>
{
    private final static IntWritable one =
        new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value,
OutputCollector<Text, IntWritable> output,
Reporter reporter) throws IOException {
        String line = value.toString();
        StringTokenizer tokenizer =
            new StringTokenizer(line);

        while (tokenizer.hasMoreTokens()){
            word.set(tokenizer.nextToken());
            output.collect(word, one);
        }
    }
}

public class WordCountReducer extends MapReduceBase
implements
Reducer<Text, IntWritable, Text, IntWritable>{

    public void reduce(Text key,
Iterator<IntWritable> values,
OutputCollector<Text, IntWritable> output,
Reporter reporter)
throws IOException{
        int sum = 0;

        while (values.hasNext())
            sum += values.next().get();

        output.collect(key, new IntWritable(sum));
    }
}
```

Fig. 5. Traditional Hadoop WordCount example.

In practice, primary map and reduce functions in Hadoop [9], write intermediate results on local disk before sending them over the network. Those I/O processes possibly imply high computing and hardware costs depending on the network-latency and disk-space costs. Thus, using combining functions minimizes the amount of intermediate data transferring from map-workers to reduce-workers. That also allow decreasing the number and size of key-value pairs to shuffle from map-workers to reduce-workers for getting improvements on the MapReduce algorithmic efficiency. Just, combining functions are named “mini-reducers”. In general, the use of combining functions seems adequate because map functions recognize intermediate-key and value pairs to send them for the shuffling and sorting process in traditional MapReduce solutions. The output of those processes corresponds to the input for reduce-worker instances.

```
private Text word = new Text();
HashMap<String, Integer> palabras =
    new HashMap<String, Integer>();

public void map(LongWritable key, Text value,
OutputCollector<Text, IntWritable> output,
Reporter reporter) throws IOException{

    palabras.clear();

    String line = value.toString();
    StringTokenizer tokenizer = new StringTokenizer(line);

    while (tokenizer.hasMoreTokens()){
        String w = tokenizer.nextToken();

        if (palabras.containsKey(w))
            palabras.put(w, palabras.get(w) + 1);

        else
            palabras.put(w, 1);
    }

    for (Entry<String, Integer> entry : palabras.entrySet()){
        String k = entry.getKey();
        Integer v = (Integer) entry.getValue();

        word.set(k);
        output.collect(word, new IntWritable(v));
    }
}
```

Fig. 6. ‘In-Mapper’ Combining Hadoop solution for the WordCount example.

A. Combining and ‘In-Mapper’ Combining

Even though map and reduce functions seem algorithmically simple to think and implement, combining function symbolize improvements performance for cases of high-traffic of data between map and reduce-workers. Combining functions act like the reduce functions [8] because they minimize the amount of intermediate data generated by each map-worker. For example, *WordCount* and *Average* represent two traditional solutions that support the use of a combining function, in the first case, functioning likes the reduce function. Nevertheless, combining functions execution are not always effective [5], [8]. Precisely, ‘In-Mapper’ Combining functions [8] solve those mentioned issues.

Fig. 5 presents a traditional Hadoop MapReduce solution for the *WordCount* example, and Fig. 6 shows an ‘In-Mapper’ Combining function to local aggregate data in the map phase and reduce the information traffic between map-worker and

```
public class Palabra {
    private String palabra;
    private Integer cuantas;

    public Palabra(String P){
        palabra = P;
        cuantas = 0;
    }

    public String getPalabra() {
        return palabra;
    }

    public void setPalabra(String palabra) {
        this.palabra = palabra;
    }

    public Integer getCuantas() {
        return cuantas;
    }

    public void incCuantas() {
        this.cuantas++;
    }
}
```

Fig. 7. Class Palabra for local aggregation in the Hadoop WordCount example.

reduce-worker. The input for that example corresponds to a set of words. Fig. 7 shows a new class *Palabra* for grouping values (local aggregation) in the *WordCount* example. The main function of the mapper functions in Fig. 5 and 6 look for identify words only, and to identify words and locally aggregate the already identified words count in the map function, respectively.

Fig. 8 shows the MapReduce solution for the *Average* example that looks for to obtain the average score of each student in a list of student and grade pairs. Fig. 9 shows an input example for the *Average* example. Fig. 10 illustrate the *GradeCount* class necessary for the local aggregation in the *Average* example.

Note that, for 'In-Mapper' Combining solution of the *WordCount* example, map function produces the same output as a traditional MapReduce solution, i.e., reduce function continues being the same. Nevertheless, as Lin and Dyer [8] illustrate, map and reduce functions of 'In-Mapper' Combining for the *Average* example do not produce and receive the same values such as those of the map and reduce functions in a traditional MapReduce solution of that example.

Even though the 'In-Mapper' Combining approach allows reducing information traffic from map-workers to reduce-workers [8], this approach implies to add more code and responsibilities on map functions. For example, 'In-Mapper' Combining of Fig. 6 includes a HashMap definition, and map function presents two actions in the loop, one to recognize each word and add them in the HashMap, and another one to update previous values of existing words; and map outputs these values after identifying all words and their occurrence number in the received input value. The number of sending and receiving messages of this solution would decrease if there were repeated words in the input. Nevertheless, map function grows in code and responsibilities, that is, the map function for 'In-Mapper' Combining approach is definitely lesser modular than its original version.

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class AverageMapper extends
Mapper<LongWritable, Text, Text, IntWritable> {
    protected void map(LongWritable key, Text value,
Context context) throws
IOException, InterruptedException {

        String line = value.toString();
        StringTokenizer tokenizer = new
StringTokenizer(line, " ");

        while(tokenizer.hasMoreElements()){
            String val = tokenizer.nextToken();

            String[] parts = val.split(" ");
            String part1 = parts[0];
            String part2 = parts[1];

            Text _id = new Text();
            _id.set(part1);
            Integer marks = new Integer(part2);

            context.write(_id, new IntWritable(marks));
        }
    }
}

import java.io.IOException;
import org.apache.hadoop.io.FloatWritable;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class AverageReducer extends
Reducer<Text, IntWritable, Text, FloatWritable>{
    protected void reduce(Text key,
Iterable<IntWritable> values,
Context context)
throws IOException, InterruptedException {
        Integer sum = 0;
        Integer cnt = 0;

        for (IntWritable value:values) {
            sum = sum + value.get();
            cnt = cnt + 1;
        }

        Float avg_m = (float) sum/cnt;

        context.write(key, new FloatWritable(avg_m));
    }
}
```

Fig. 8. Traditional Hadoop Average example.

```
Valery 6, Maria 4,
Paul 3, Paul 7,
Patrick 7, Fabrice 10,
Valery 10, Chris 0,
Fabrice 0, Chris 10,
Paul 10, Patrick 9,
Valery 2, Patrick 10,
Patrick 0, Sebas 9
```

Fig. 9. Input format for the Hadoop Average example.

```
import java.io.DataInput;
import java.io.DataOutput;
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Writable;

public class GradeCount
    implements Writable {
    private IntWritable grade;
    private IntWritable count;

    public GradeCount() {
        this.grade = new IntWritable();
        this.count = new IntWritable();
    }

    //Custom Constructor
    public GradeCount(IntWritable grade,
        IntWritable count) {

        this.grade = grade;
        this.count = count;
    }

    //Setter method to set the values
    //of a GradeCount object
    public void set(IntWritable grade,
        IntWritable count) {

        this.grade = grade;
        this.count = count;
    }

    //to get Count from GradeCount Record
    public IntWritable getCount()
    {
        return count;
    }

    @Override
    //overriding default readFields method.
    //It de-serializes the byte stream data
    public void readFields(DataInput in)
        throws IOException {

        grade.readFields(in);
        count.readFields(in);
    }

    @Override
    //It serializes object data into
    //byte stream data
    public void write(DataOutput out)
        throws IOException {

        grade.write(out);
        count.write(out);
    }
}
```

Fig. 10. Class GradeCount for local aggregation in the Hadoop Average example.

B. Aspect-Combining

Aspect-Combining represents a combining function as an AOP aspect on map function. In practice, such as Fig. 11 illustrate, AOP solutions would permit add behavior on MapReduce map methods just to isolate their functioning and nature. Thus, we propose Aspect-Combining. Aspect-Combining looks for the inclusion of structural and functioning elements of traditional ‘In-Mapper’ Combining functions. Hence, Aspect-Combining preserves the simplicity of the map function and guarantees the execution of the function combining. Furthermore, Aspect-Combining seems applicable by the use of any

AOP approach over Hadoop. Next, this article describes a few AspectJ application examples.

Like for traditional combining function, the goal of Aspect-Combining is to locally aggregate data in map-worker instances to diminish the associated networking traffic in the map-workers for the shuffling process. Therefore, taking into account the components and functioning of the ‘In-Mapper’ Combining solutions such as those in Fig. 6; a class that contains the map function should also contain an attribute for local aggregation and methods for that process. Thus, in the *WordCount* example, it is necessary to know about each identified word and the number of previous occurrences of that word for updating its occurrences number. Hence, new attributes and methods for advisable classes are required by inter-type declaration in an AOP context. Likewise, in *Average* case, for each identified student, it would be necessary to sum their grades and also to count the number of their grades.

As Fig. 11 shows, three events exist for code injection in the advisable map method: before starting the execution of a map method to initialize attributes to group values, around the execution of a map method to group or create an identified element for local aggregation, and after the method map finishes for sending information to the next MapReduce step. Without considering the injection time for the occurrence of these events, pointcut rules are definable in AOP and AspectJ as well as the time for injecting the new behavior code that is analogue to the definition for AOP advices.

IV. ASPECT-COMBINING APPLICATION AND RESULTS

A. Experiments

Table I shows the hypothesis and use of variables when conducting experimentation on classic Combining, In-Mapper Combining, and Aspect-Combining functions on the *Word-Count* and *Average* examples.

For each experiment of Table I, the null hypothesis establishes that Aspect-Combining neither performs faster nor is more modular than classic Combining and ‘In-Mapper’ Combining solutions on the analyzed examples.

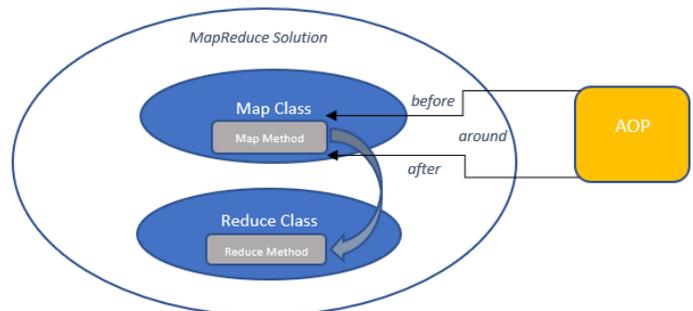


Fig. 11. Advisable map method of a MapReduce solution.

B. Results

In this section we discuss the results we obtained and how the null hypothesis has been rejected, thus accepting the alternative hypothesis.

We perform four experiments on the *WordCount* example and three experiments on the *Average* example to check the validity of the Aspect-Combining approach for modular MapReduce solutions in Hadoop.

Fig. 12 and 13 present the definition of pointcut instances for the Aspect-Combining of the *WordCount* and *Average* examples, in this case, 3 point cuts for each case: one to start collecting data, one for the collect and write methods call inside the advised map methods, and one for the end of the map method execution.

Fig. 14 and 15 show Aspect-Combining inter-type declaration for the *WordCount* *Average* examples to add and manipulate the required object collection, *ArrayList* of class *Palabra* and *HashMap* of class *GradeCount* instances, respectively.

Finally, for Aspect-Combining in the *WordCount* and *Average* examples, Fig. 16 and 17 present advice instances for before the method map execution to initialize the attribute for local aggregation, around the grouping values process, and after the execution of map method to effectively send the locally grouped values to the next MapReduce step.

As a practical functioning and results evaluation, Tables II and III present traditional In-Mapper and Aspect-Combiner results for the *WordCount* and *Average* examples to appreciate and compare them. We run practical experiments in a single Lenovo ThinkPad Edge E530 laptop of 2.50 GHz, 16GB of RAM and a Core i3 processor. For the *WordCount* examples, as input files, *Words* is a text file of 168 bytes, and *ebook* is a file of 1.6 MB; whereas for the *Average* examples, input files were generated taking in account ten students, and grades from 0 to 100.

Although these experiments did not run in a cluster of computing machines, and knowing the main practical improvement of 'In-Mapper' and Aspect-Combining is a reduction of traffic between map-workers and reduce-workers; surprisingly, Aspect-Combining permits obtaining better modularity and better performance for big-input examples. Hence, only for a single and small file, the traditional *WordCount* solution without combining approaches obtains the best time. In the *WordCount* example, for two files, 'In-Mapper' *WordCount* solution is the best and Aspect-Combining the 2nd one. For all other cases, Aspect-Combining presents the best performance. Thus, in addition to the best modularity, Aspect-Combining permits getting efficient computation results in one machine execution. This situation would be the same cluster environments.

V. DISCUSSION

The null hypothesis establishes that Aspect-Combining does not improve the modularity for the presence of crosscutting-concerns issues and the execution-time compared to the Combining and 'In-Mapper' Combining solutions for testing on the *WordCount* and *Average* examples. To refute that hypothesis, we review the modular code of Aspect-Combining solutions in which the map function has only one responsibility, and we analyze the execution-time for experiments described in Table II and Table III, both tables for random files of different sizes. For the appreciated results, we accepted the alternative hypothesis that the Aspect-Combining

```
pointcut init(WordCountMapper mapper):  
    execution(* map(..)) && target(mapper);  
  
pointcut send(WordCountMapper mapper, Text word,  
    IntWritable val):  
    call(* collect(..)) && args(word, val) &&  
    this(mapper)  
    && !within(AspectMapper);  
  
pointcut end(WordCountMapper mapper, LongWritable key,  
    Text value, OutputCollector<Text, IntWritable> output,  
    Reporter reporter): execution(* map(..)) &&  
    args(key, value, output, reporter) &&  
    target(mapper);
```

Fig. 12. Pointcut definition for Aspect-Combining in the *WordCount* example.

```
aspect AspectMapper {  
    pointcut init(AverageMapper mapper):  
        execution(* map(..)) && target(mapper);  
  
    pointcut send(AverageMapper mapper, Text st,  
        GradeCount gc):  
        call(* write(..)) && args(st, gc) &&  
        this(mapper)  
        && !within(AspectMapper);  
  
    pointcut end(AverageMapper mapper, LongWritable key,  
        Text value, Context context): execution(* map(..)) &&  
        args(key, value, context) &&  
        target(mapper);
```

Fig. 13. Pointcut definition for Aspect-Combining in the *Average* example.

outperforms the traditional approach of Combining and 'In-Mapper' Combining solutions on the *WordCount* and *Average* case-studies.

The SRP establishes that each module or class should have one and only one purpose and reason to change since if a class has more than one responsibility, then the responsibilities become coupled [10]. According to [7], "The SRP is the OOD solution to the classic 'separation-of concerns' problem". Thus, Aspect-Combining permits simple map functions and efficient MapReduce solutions, even though, for the weaving process of AOP, the code of 'In-Mapper' Combining and the final one of Aspect-Combining should be equivalent.

```
private ArrayList<Palabra> WordCountMapper.palabras;  
  
public void WordCountMapper.initPalabras(){  
    palabras = new ArrayList<Palabra>();  
}  
  
public void WordCountMapper.incPalabra(String palabra){  
    if (palabras.contains(palabra)){  
        Integer index = palabras.indexOf(palabra);  
        Palabra p = palabras.get(index);  
        p.incCuantas();  
        palabras.set(index, p);  
    }  
    else  
        palabras.add(new Palabra(palabra));  
}  
  
public ArrayList<Palabra> WordCountMapper.getPalabras(){  
    return palabras;  
}
```

Fig. 14. Inter-type declaration for Aspect-Combining in the *WordCount* example.

TABLE I. HYPOTHESES AND DESIGN OF EXPERIMENTS FOR WORDCOUNT AND AVERAGE MAPREDUCE EXAMPLES

Hypotheses of Experiments 1 and 2	
Null Hypothesis (H_0)	Aspect-Combining solutions neither are faster nor more modular than Combining and 'In-Mapper' Combining for the <i>WordCount</i> case-study.
Alt. Hypothesis (H_1)	Exist cases in which Aspect-Combining solutions performs faster than Combining and In-Mapper Combining and does not present crosscutting concerns for the <i>WordCount</i> case-study.
files used as input	Randomly generated files of words of name and grade. Size of files are from 10KB to 1MB.
Blocking variables	In each experiment, we generated a set of files in increasing size.
Hypotheses of Experiment 2	
Null Hypothesis (H_0)	Aspect-Combining solutions neither are faster nor more modular than Combining and 'In-Mapper' Combining for the <i>Average</i> case-study.
Working Hypothesis (H_1)	Exist cases in which Aspect-Combining solutions perform faster than Combining and 'In-Mapper' Combining, and Aspect-Combining solutions do not present crosscutting concerns for the <i>Average</i> case-study.
Files used as input	Randomly generated files of pairs of name and grade. Size of files are from 10KB to 1MB.
Blocking variables	In each experiment, we generated a set of files in increasing size.
Constants	
Hadoop 2.4.1 in Ubuntu Linux 14.02	<i>WordCount</i> and <i>Average</i> solutions implemented in 2016 and 2017, respectively

TABLE II. WORDCOUNT SOLUTIONS - PRACTICAL EVALUATION

Input	Traditional WordCount	In-Mapper WordCount	Aspect-Combiner WordCount
Words file (168B)	2768867581 ns	2797626514 ns	2830724641 ns
Words file (168B)+ ebook (1.6MB)	6474883019 ns	4750820118 ns	5675306443 ns
Words file (168B) + 30 ebook copies (48MB)	36835481761 ns	36011355288 ns	33913147695 ns
Words file (168B) + 50 ebook copies (80MB)	52135395542 ns	58132499534 ns	51053756385 ns

TABLE III. AVERAGE SOLUTIONS - PRACTICAL EVALUATION

Input	Traditional Average	In-Mapper Average	Aspect-Combiner Average
100 files (48.1KB)	12208547623 ns	11833383768 ns	13055207744 ns
200 files (192.6MB)	139049238451 ns	61734379340 ns	68603119714 ns
400 files (384.1MB)	257664874145 ns	134202559848 ns	129013569364 ns

```
private Map<String, GradeCount> AverageMapper.scores;

public void AverageMapper.initScores(){
    scores = new HashMap<String, GradeCount>();
}

public void AverageMapper.addScore(String st, GradeCount gc){
    if (scores.containsKey(st)) {
        GradeCount ss = (GradeCount) scores.get(st);

        Integer sum = Integer.parseInt(ss.getGrade().toString()) +
            Integer.parseInt(gc.getGrade().toString());
        Integer count = Integer.parseInt(ss.getCount().toString()) +
            gc.getCount();
        ss.set(new IntWritable(sum), new IntWritable(count));
    } else {
        scores.put(st, gc);
    }
}

public Map<String, GradeCount> AverageMapper.getScores(){
    return scores;
}

before(WordCountMapper mapper): init(mapper){
    mapper.initPalabras();
}

void around(WordCountMapper mapper, Text word,
    IntWritable val): send(mapper, word, val){
    mapper.incPalabra(word.toString());
}

after(WordCountMapper mapper, LongWritable key, Text value,
    OutputCollector<Text, IntWritable> output,
    Reporter reporter) throws IOException:
    end(mapper, key, value, output, reporter){

    ArrayList<Palabra> palabras = mapper.getPalabras();

    for(Palabra p: palabras){
        Text w = new Text();
        IntWritable c = new IntWritable();

        w.set(p.getPalabra());
        c.set(p.getCuantas());

        output.collect(w, c);
    }
}
```

Fig. 15. Inter-type declaration for Aspect-Combining in the Average example.

VI. CONCLUSIONS

In this section, we present the lessons we learned while developing the Aspect-Combining solutions:

- Aspect-Combining presents a practical symbiosis between MapReduce and AOP. In particular, this article presented a Hadoop and AspectJ for the implementation of Aspect-Combining.

Fig. 16. Advices for Aspect-Combiner in the WordCount example.

```
before(AverageMapper mapper): init(mapper){
    mapper.initScores();
}
void around(AverageMapper mapper, Text st,
    GradeCount val): send(mapper, st, val){
    mapper.addScore(st.toString(), val);
}
after(AverageMapper mapper, LongWritable key,
    Text value, Context context)
    throws IOException, InterruptedException:
end(mapper, key, value, context){
    Map<String, GradeCount> sts = mapper.getScores();
    Iterator<Map.Entry<String, GradeCount>> itr1 =
        sts.entrySet().iterator();
    while (itr1.hasNext()) {
        Entry<String, GradeCount> entry1 =
            itr1.next();
        String s_id_1 = entry1.getKey();
        Text t = new Text(); t.set(s_id_1);
        GradeCount ss = entry1.getValue();
        context.write(t, ss);
    }
}
```

Fig. 17. Advices for Aspect-Combiner in the Average example.

- Thinking on the primary functions of MapReduce along with their focus, original combining functions are usually adequate to preserve the map function nature and simplicity. Nonetheless, this article pointed out its non-effectiveness and cost. Therefore, ‘In- Mapper Combining seems more practical, but they do not respect modularity principles. Hence, this article presented and practically proved the benefits of Aspect-Combining for modular MapReduce solutions and, for big data-input, possible more efficient results than Combining and ‘In-Mapper’ Combining Hadoop solution.
- Although a class for map-worker permit the production of modular solutions, a programmer is in charge of putting attention on Initialize, Map, and Close methods, that is, setup(..), map(..), and cleanup(..) methods in Hadoop which does not permit an independent development. Thus, Aspect-Combining approach separates these functions as advice instances, and the map-worker focuses only on an oblivious map(..) method of before(..), around(..) and after(..) advice instances which operate similar to Initialize, Map, and Close methods of Fig. 18 [8].

```
class MAPPER
    method INITIALIZE
        H ← new ASSOCIATIVEARRAY
    method MAP(docid a, doc d)
        for all term t ∈ doc d do
            H{t} ← H{t} + 1
    method CLOSE
        for all term t ∈ H do
            EMIT(term t, count H{t})
```

Fig. 18. A modular structure of Mapper class in MapReduce Solutions.

As future work, this research group plans to review more about AOP on MapReduce applications to figure out the applicability of other AOP practical approaches such as JPI [18], [20], [22], [24] and Ptolemy [17] on Hadoop [5], [8] and Giraph approaches [25], and compare their effectiveness and practical performance. Giraph also permits defining combining functions without a guarantee for their execution [25], and Aspect-Combining seems adequate to guarantee their execution.

ACKNOWLEDGMENT

This work was partially supported by CONICYT-CCV/Doctorado Nacional/2018-21181055.

REFERENCES

- [1] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” in *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation - Volume 6*, ser. OSDI’04. Berkeley, CA, USA: USENIX Association, 2004, pp. 10–10. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1251254.1251264>
- [2] —, “Mapreduce: Simplified data processing on large clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008. [Online]. Available: <http://doi.acm.org/10.1145/1327452.1327492>
- [3] —, “Mapreduce: A flexible data processing tool,” *Commun. ACM*, vol. 53, no. 1, pp. 72–77, Jan. 2010. [Online]. Available: <http://doi.acm.org/10.1145/1629175.1629198>
- [4] D. Miner and A. Shook, *MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems*, 1st ed. O’Reilly Media, Inc., 2012.
- [5] T. White, *Hadoop: The Definitive Guide*, 4th ed. O’Reilly Media, Inc., 2015.
- [6] G. Kiczales, “Aspect-oriented Programming,” *ACM Comput. Surv.*, vol. 28, no. 4es, dec 1996. [Online]. Available: <http://doi.acm.org/10.1145/242224.242420>
- [7] D. Wampler, “Aspect-oriented design principles: Lessons from object-oriented design,” *Proceedings of the Sixth International Conference on Aspect-Oriented Software Development*, vol. AOSD’07, pp. 615–636, 2007.
- [8] J. Lin and C. Dyer, *Data-Intensive Text Processing with MapReduce*. Morgan and Claypool Publishers, 2010.
- [9] J. A. Galindo, M. Acher, J. M. Tirado, C. Vidal, B. Baudry, and D. Benavides, “Exploiting the enumeration of all feature model configurations: A new perspective with distributed computing,” in *Proceedings of the 20th International Systems and Software Product Line Conference*, ser. SPLC ’16. New York, NY, USA: ACM, 2016, pp. 74–78. [Online]. Available: <http://doi.acm.org/10.1145/2934466.2934478>
- [10] R. C. Martin, *Agile Software Development: Principles, Patterns, and Practices*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2003.
- [11] R. Laddad, *AspectJ in Action: Enterprise AOP with Spring Applications*, 2nd ed. Greenwich, CT, USA: Manning Publications Co., 2009.
- [12] R. Miles, *AspectJ Cookbook*. O’Reilly Media, Inc., 2004.
- [13] J. D. Gradecki and N. Lesiecki, *Mastering AspectJ: Aspect-Oriented Programming in Java*. New York, NY, USA: John Wiley & Sons, Inc., 2003.
- [14] S. Apel, D. Batory, C. Kstner, and G. Saake, *Feature-Oriented Software Product Lines: Concepts and Implementation*. Springer Publishing Company, Incorporated, 2013.
- [15] S. Apel, D. Batory, and M. Rosenmüller, “On the Structure of Cross-cutting Concerns: Using Aspects or Collaborations?” Oct. 2006.
- [16] G. Kiczales and M. Mezini, “Aspect-oriented programming and modular reasoning,” in *Proceedings of the 27th International Conference on Software Engineering*, ser. ICSE ’05. New York, NY, USA: ACM, 2005, pp. 49–58. [Online]. Available: <http://doi.acm.org/10.1145/1062455.1062482>

- [17] H. Rajan, G. T. Leavens, R. Dyer, and M. Bagherzadeh, "Modularizing crosscutting concerns with ptolemy," in *Proceedings of the Tenth International Conference on Aspect-oriented Software Development Companion*, ser. AOSD '11. New York, NY, USA: ACM, 2011, pp. 61–62. [Online]. Available: <http://doi.acm.org/10.1145/1960314.1960332>
- [18] E. Bodden, E. Tanter, and M. Inostroza, "Join Point Interfaces for Safe and Flexible Decoupling of Aspects," *ACM Trans. Softw. Eng. Methodol.*, vol. 23, no. 1, pp. 7:1–7:41, Feb. 2014. [Online]. Available: <http://doi.acm.org/10.1145/2559933>
- [19] C. Vidal Silva, R. Villarroel, R. Schmal, R. Saens, C. Del Rio, and T. Tigero, "Aspect-Oriented Formal Modeling: (AspectZ + Object-Z) = OOAspectZ," *COMPUTING AND INFORMATICS*, vol. 34, no. 5/15, 2015.
- [20] C. Vidal Silva, R. Villarroel, L. López, M. Bustamante, R. Schmal, and V. Rea, "JPI UML Software Modeling: Aspect-Oriented Modeling for Modular Software," *International Journal of Advanced Computer Science and Application IJACSA*, vol. 6, no. 12, 2015.
- [21] E. Bodden, "Closure Joinpoints: Block Joinpoints Without Surprises," ser. AOSD '11. New York, NY, USA: ACM, 2011, pp. 117–128. [Online]. Available: <http://doi.acm.org/10.1145/1960275.1960291>
- [22] C. Vidal Silva, R. Saens, C. Del Rio, and R. Villarroel, "Aspect-Oriented Modeling: Applying Aspect-Oriented UML Use Cases and Extending Aspect-Z," *COMPUTING AND INFORMATICS*, vol. 32, no. 3, 2013.
- [23] —, "OOAspectZ and aspect-oriented UML class diagrams for Aspect-oriented software modelling (AOSM)," *INGENIERIA E INVESTIGACION*, vol. 33, no. 3, 2013.
- [24] C. Vidal Silva, R. Villarroel, and C. Pereira, "JPIAspectZ: A formal specification language for Aspect-Oriented JPI applications," *Proceedings of XXXIII International Conference of the Chilean Computer Science Society*, 2014.
- [25] R. Shaposhnik, C. Martella, and D. Logothetis, *Practical Graph Analytics with Apache Giraph*, 1st ed. Berkely, CA, USA: Apress, 2015.

A New Message Encryption Method based on Amino Acid Sequences and Genetic Codes

Ahmed Mahdee Abdo
Computer Science Department
Faculty of Science
University of Zakho
Zakho, Kurdistan, Iraq

Adel Sabry Essa, Abdullah A. Abdullah
Computer Science Department
Faculty of Science
University of Zakho
Duhok, Kurdistan, Iraq

Abstract—As the use of technology is increasing rapidly, the amount of shared, sent, and received information is also increasing in the same way. As a result, this necessitates the need for finding techniques that can save and secure the information over the net. There are many methods that have been used to protect the information such as hiding information and encryption. In this study, we propose a new encryption method making use of amino acid and DNA sequences. In addition, several criteria including data size, key size and the probability of cracking are used to evaluate the proposed method. The results show that the performance of the proposed method is better than many common encryption methods, such as RSA in terms of evaluation criteria.

Keywords—Information; secure; encryption

I. INTRODUCTION

As the digital world is growing dramatically, the need for secure and safe information is growing in the same way. There are many ways to keep the data secure such as Cryptography, Stenography or a combination of them [1]. There are two main ways to encrypt a text, either secret-key cryptography or public-key cryptography [2]. The author in [3] proposed an algorithm to encrypt a message based on DNA sequences. The algorithm uses a DNA sequence to encrypt a text using one of the complementary rules. He proposed three complementary rules based on biological and chemical features of each DNA base among each other. The algorithm also made a DNA based code to represent each DNA letter with 2 bits binary number. The final string contains English alphabetic letters that involve the hidden message. The author in [4] use another method to encrypt and hide the data in a DNA sequence. The technique uses a dictionary of codons that is used in our own method. The dictionary codon contains on 64 codons starting from ACT which is numbered 0 to GGG that is numbered 111111. The method converts the message to 8 bits ASCII binary number. Then each 6 bits of the message are represented to three DNA letter based on dictionary codon. In our new proposed algorithm, a new technique is being used in the process of cryptographic taking benefits of amino acid sequences. The rest of the paper will explain the details of the new method. Firstly, it presents a background about the information security. Then it will explore basic of molecular biology. Thereafter, the steps of the encryption and decryption with an example will be explained. Finally, the paper will assess the new method based on some criteria with existing methods.

II. INFORMATION SECURITY

There are millions of people who get in touch to each other every day electronically via e-mail, e-commerce, e-banking machine and e-learning [2]. Among these vast amount of communications over the net, the biggest question is to which extend these communications are secure. Transferring information across the world over the Internet leads to the biggest concern related to the security of the transferred information. There are many attacks, such as Crypt analytic and Brute force, could recover the original message when the information are sensitive. Encryption is one of the methods used to secure the information. It has been developed quickly recent years to save and protect transmitted information [5].

A. Encryption

Encryption is a mathematical technique which is related to the security of information sides such as data authentication, data integrity and confidentiality. Encryption is used wildly in many fields especially in sensitive communications such as in wars, military bases and intelligent agency. Cryptography is not just meant to information security, it is rather set of techniques. There are two types of Cryptography which are used widely, public and secret key. Rivest Cipher 5 (RC5) and Advanced Encryption Standard (AES) are the most well-known algorithms based on secret key, while RivestShamirAdleman (RSA) algorithm is the most common for public key [2].

III. BASICS OF MOLECULAR BIOLOGY

The life of an organism has been mapped as a very long sequence called genome. The genome is a repeated of 4 chemical bases called deoxy ribo nucleic acid (DNA). The DNA consists of 4 nucleotides (A, C, G and T). The ribo nucleic acid (RNA) sequence is one of genetic materials. The DNA sequence is convert into mRNA sequence in an operation called transaction. Then the mRNA is translated into amino acid sequence based on genetic codes. There are 20 amino acid letters which construct any protein [6]. The twenty amino acids are (E, P, A, C, G, Q, V, R, K, W, D, N, H, F, L, I, S, T, Y, M).

A. BLUSOM 50

BLOSOM (Blocks Substitution Matrix) are substitution matrices used in Bioinformatics for aligning amino acid sequences to give a score of each alignment. The intersection of

amino acids in the matrix is a specific score which represents to what extent it is willing to interact with other amino acids in the matrix. The higher score in the matrix is when the amino acid interacts with itself [7]. BLUSOM50 is a scoring matrix that is used by FASTA and BLAST programs for identifying distant homologous especially with fully length sequence [8].

IV. PROPOSED APPROACH

Based on the score among of the amino acids in BLUSOM 50, four amino acids which have less score with other amino acids have been neglected from the proposed algorithm. The amino acids that are not considered are (S, T, Y, M) while the rest sixteen amino acids (E, P, A, C, G, Q, V, R, K, W, D, N, H, F, L, I) have been used. Then each of these amino acids has been given a 4 bits binary number as it is illustrated in Table 1.

TABLE I. REPRESENTING EACH AMINO ACID BY 4 BITS BINARY NUMBER

Amino acid	Code	Amino acid	Code
A	0000	C	0100
R	0001	Q	0101
N	0010	E	0110
D	0011	G	0111
H	1000	F	1100
I	1001	P	1101
L	1010	W	1110
K	1011	V	1111

Two kinds of complementary rules have been applied. A complementary rule means which amino acid is more or less applicable to interact with other amino acids according to the scores taken from BLOSUM50. The higher the score, the higher the probability of interaction. The lower the score, the lower the interaction probability. Accordingly, the first complementary rule is based on the maximum score between a specific amino acid and the rest, and then we take that amino acid as a complementary rule for the proposed approach. The second complementary rule is based on the minimum score between a specific amino acid and the rest, and then we take that amino acid as a complementary as it is shown in Tables 2 and 3.

TABLE II. RULE 1, THE COMPLEMENTARY RULE ACCORDING TO MAXIMUM VALUE IN BLOSUM50

(A, G), (G, A)	(R, Q),(Q, R)	(N, D), (D, N)	(C, V), (V, C)
(E, K), (K, E)	(H, F), (F, H)	(I, L), (L, I)	(P, W), (W, P)

TABLE III. RULE 2, THE COMPLEMENTARY RULE ACCORDING TO MINIMUM VALUE IN BLOSUM50

(A, F), (F, A)	(R, C),(C, R)	(N, L), (L, N)	(K, P), (P, k)
(Q, D), (I, Q)	(E, G), (G, E)	(H, V), (V, H)	(D, W), (W, D)

In addition, a 6 bits binary number has been given for every codon starting from 000000 for TTT to 111111 for GGG as it is revealed in Table 3.

A. Proposed Data Encryption Algorithm: Main Steps

- 1) Represent each letter in the original message as a binary of 8 bits.
- 2) Represent each part (4 bits) as a letter based on Table 1.

TABLE IV. DICTIONARY OF CODONS, ADOPTED [4]

Codon	6 Bits number	Codon	6 Bits number
TTT	000000	TTC	010000
TCT	000001	TCC	010001
TAT	000010	TAC	010010
TGT	000011	TGC	010011
CTT	000100	CTC	010100
CCT	000101	CCC	010101
CAT	000110	CAC	010110
CGT	000111	CGC	010111
ATT	001000	ATC	011000
ACT	001001	ACC	011001
AAT	001010	AAC	011010
AGT	001011	AGC	011011
GTT	001100	GTC	011100
GCT	001101	GCC	011101
GAT	001110	GAC	011110
GGT	001111	GGC	011111
TTA	100000	TTG	110000
TCA	100001	TCG	110001
TAA	100010	TAG	110010
TGA	100011	TGG	110011
CTA	100100	CTG	110100
CCA	100101	CCG	110101
CAA	100110	CAG	110110
CGA	100111	CGG	110111
ATA	101000	ATG	111000
ACA	101001	ACG	111001
AAA	101010	AAG	111010
AGA	101011	AGG	111011
GTA	101100	GTG	111100
GCA	101101	GCG	111101
GAA	101110	GAG	111110
GGA	101111	GGG	111111

- 3) Apply one of the complementary rule (either rule one or two) on the letter) based on Table 2.
- 4) Take the first occurrence position of the 16 amino acids in any sequence as indexes.
- 5) Represent the decimal index numbers to 8 bits binary numbers.
- 6) Finally, represent each 6 bits binary numbers represented to DNA letters according to Table 3. Extra zeros are added to make the length of binary number 6 or multiply of 6.

B. An Example to Encrypt a Message

Let consider our message is Hi, and the amino acid sequence is MPQVKLWLSGIQICLQSNQLAPLIRELQKD-STASFHFI EGEVECGPGPGIEGIFEGP

- 1) Represent the message in 8bits binary: 01001000, 01101001
- 2) Represent each part (4 bits) as a letter based on Table 1: C, H, E, I
- 3) Apply one of the complementary rule (1) based on Table 2: V, F, K, L
- 4) The first occurrence of V F K L in the sequence is 3, 34, 4, 5 respectively.
- 5) Represent the decimal index numbers to 8 bits binary: 00000011, 00100010 00000100, 00000101
- 6) Represent each 6 bits binary numbers represented to DNA letters according to Table 3. Not: extra zero added to make all of them 6bits 000000 = TTT, 110010 =TAG, 001000 = ATT, 000100 = CTT, 000001= TCT, 010000 = TTC The faked DNA that hold the secret message is TTTTA-GATTCTTTCTTTTC

C. Proposed Data Decryption Algorithm: Main Steps

- 1) Represent each three DNA letters of the encrypted message to 6 bits binary numbers according to Table 3.
- 2) Take each 8bits and then convert them decimal numbers. Ignore the mod of 8.
- 3) Take these decimal numbers as indexes for the key and retrieve the represented value for each index in the key.
- 4) Apply one of the complementary rule (either rule one or two) on the letter created in previous step, so each letter will be represented to another letter of amino acids based on Table 2.
- 5) Represent each letter to 4bits binary number according to Table 1.
- 6) Represent each 8bits binary number to ASCII numbers and letters to get your original message.

D. An Example to Decrypt a Message

The faked DNA is: TTTTAGATTCTTTCTTTC
The amino acid sequence: MPQVKLWLSGIQICLQS-
NQLAPLIRELQKD STASFHFIEGEVECGPGGIEGIFEGP

- 1) Represent each three DNA letters of the faked DNA to 6 bits binary numbers according to Table 3: TTT = 000000, TAG= 110010, ATT= 001000, CTT= 000100, TCT= 000001, TTC= 010000.
- 2) Represent every 8bits to decimal number: 3, 34, 4, 5 Ignore the mode of 8 (here is the last 4 zeros).
- 3) Take these decimal numbers as indexes for the key and retrieve the represented value for each index in the key: V, F, K, L.
- 4) Apply complementary rule (1): C, H, E, I
- 5) Represent each letter to 4bits binary number according to Table 1: 0100, 1000, 0110, 1001
- 6) Represent each 8bits binary number to ASCII numbers and letters to get your original message: 01001000 = H 01101001 = i The message is Hi

V. RESULTS AND COMPARATIVE ANALYSIS

The results show high performances in the proposed method comparing with other encryption techniques. There are many criteria that can be used to evaluate any encryption method or algorithm. These terms are key size, data size, security and time complexity [9]. Here, we present a comparative analysis between our proposed algorithm and well-known encryption method based on these criteria.

A. Key Size

Key size means the length of the key used in bits or letters to encrypt a message by any technique. Every method of encryption has its own key size [10]. In the asymmetric RSA cipher, the amount of key size is not limited to any number. It could be any length as it depends on the amount of information to be encrypted. The more the information, the longer the key, the more complexity time. AES is another method which has three options in terms of key size, either 256 or 192 or 128. This increase the security level of this technique as the key size has a positive correlation with security level

[1]. In our proposed algorithm, the key size is composed of 16 amino acid letters. These letters are extracted from any amino acid sequence with their indexes to be used for the encryption process. The amino acid sequence could be in any length and the 16 letter could be in any position. Consequently, this increase the security level of this method to a large extend as there are hundreds of thousands of amino acid sequences.

B. Data Size

The amount of data to be encrypted which is measured by either bytes or kilobytes is called data size. The data size in some cases could be a weak point for some encryption methods. Some methods has restriction size on the amount of data to be encrypted. While others such as vigenere cipher method, the big amount of data led to repetition when the key size is small. In RSA cipher case, there is a positive relation between the key size and the data size. Each amount of data should has an appropriate key size [1]. In this new method, the data size has no limitation. In addition, it has no relation with key size. Accordingly, each of the key and the data are separated from each other.

C. Security

The security in the field of encryption means the probability of cracking. Every encryption technique has its own security level by calculating the cracking probability. The security level of the Vigenere cipher method is very weak as the key is repeated, so the cracker can guess the key length of the key using some statistical methods and ultimately decipher the text. On the other hand, the RSA cipher security level relies heavily on the key size. If the chosen key is small, the text could be cracked easily. In the AES method, the time required to crack it, depends on the key length used to cipher the text. As AES use long key size, it makes its method better than other advanced symmetric ciphers and cannot be cracked only by using the brute force cryptanalysis. In our proposed method, the key is an amino acid sequence in any length. The indexes of the first occurrence of the 16 amino acids are kept to be used later. Accordingly, the security level of the new encryption method depends on how many amino acid sequences are there?. There are many official databases that hold amino acid sequences. The Universal Protein Resource (UniProt) is one of the main amino acid and protein sequences recourse over the world. According to last release on March 2017, the database of the UniProt contains on 80204459 amino acid sequences [11]. So the security level or the probability of cracking can be calculated as it is illustrated below. The probability of cracking will be $=1/280204459$. We multiplied it by 2 as we use 2 complementary rules. In addition, this is not the end, this number (80204459) is grown dramatically and this will increase the security level of the method

D. Time Complexity

Time complexity means the required time needed to encrypt a message. The time depends heavily on the steps of a specific algorithm. In our algorithm, there are 6 steps to encrypt a message. The more advanced ciphering methods such as AES and RSA shows more executing time than simpler methods such as DNA-based PCF.

We used Perl programming language to implement our method. gmtime() is used in Perl to print the executing time. It shows very fast performance to encrypt even a long message.

All the criteria above have been compared and elaborated in Table 5 which compare the proposed method with common once

TABLE V. COMPARATIVE ANALYSIS

Parameters	DNA-based PCF	RSA	Proposed method
Key size	English letters of maximum length 25	Large sizes lead to extensive computational efforts	16 amino acids letters.
Data size	Any size	Large sizes needs large key size	Any size
Time complexity	Low	The highest	Low
Security	Its cracking probability is very low	Can be cracked for small key lengths	Very low and explained in security point.

VI. CONCLUSION

To conclude, it is compulsory to keep the information safe. Using DNA and amino acids sequences as mediums to encrypt a message took much attention of the researchers. There are many algorithms and methods have been designed to encrypt a message using different kind of medium. Every cryptography technique has its own features in term of performance. Some of them are performing well in terms of key size and might fail in data size, while other present good result in time complexity meanwhile has high probability to crack the message.

Our proposed method has no restrictions in data size and the key size. In addition, its probability to crack it down is very low comparing with other methods. Moreover, this probability is decreased as long as UniProt get updated. In terms of time, as there are just six steps to encrypt any message, it shows high speed in terms of time complexity. Consequently, the new proposed algorithm has satisfied most encryption criteria to a

large extend the used to evaluate and ciphering method. This method could be more improved by decreases the length of the encrypted message.

ACKNOWLEDGMENT

Thanks to everyone helped me to finish this work.

REFERENCES

- [1] S. Marwan, A. Shawish, and K. Nagaty, "Dna-based cryptographic methods for data hiding in dna media," *Biosystems*, vol. 150, pp. 110–118, 2016.
- [2] V. Kamalakannan and S. Tamilselvan, "Security enhancement of text message based on matrix approach using elliptical curve cryptosystem," *Procedia Materials Science*, vol. 10, pp. 489–496, 2015.
- [3] K. Menaka, "Message encryption using dna sequences," in *Computing and Communication Technologies (WCCCT), 2014 World Congress on. IEEE*, 2014, pp. 182–184.
- [4] R. Agrawal, M. Srivastava, and A. Sharma, "Data hiding using dictionary based substitution method in dna sequences," in *Industrial and Information Systems (ICIIS), 2014 9th International Conference on. IEEE*, 2014, pp. 1–6.
- [5] A. Joshi, M. Wazid, and R. Goudar, "An efficient cryptographic scheme for text message protection against brute force and cryptanalytic attacks," *Procedia Computer Science*, vol. 48, pp. 360–366, 2015.
- [6] V. Mathura and P. Kanguane, *Bioinformatics: a concept-based introduction*. Springer Science & Business Media, 2008.
- [7] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10915–10919, 1992.
- [8] W. R. Pearson, "Selecting the right similarity-scoring matrix," *Current protocols in bioinformatics*, vol. 43, no. 1, pp. 3–5, 2013.
- [9] D. A. A. G. Singh and R. Priyadharshini, "Performance analysis of data encryption algorithms for secure data transmission," *International Journal for Science and Advance Research In Technology*, vol. 2, no. 12, 2016.
- [10] W. Stallings, *Cryptography and network security: principles and practice*. Pearson Upper Saddle River, NJ, 2017.
- [11] U. Consortium, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2016.

Using Artificial Neural Networks for Detecting Damage on Tobacco Leaves Caused by Blue Mold

Himer Avila-George*, Topacio Valdez-Morones†, Humberto Pérez-Espinosa‡, Brenda Acevedo-Juárez*, and Wilson Castro§¶

*Centro Universitario de los Valles, Universidad de Guadalajara. Ameca, Jalisco 46600, México

†Unidad de Transferencia Tecnológica Tepic. Tepic, Nayarit 63173, México

‡CONACYT - CICESE, Unidad de Transferencia Tecnológica Tepic. Tepic, Nayarit 63173, México

§Facultad de Ingeniería, Universidad Privada del Norte. Cajamarca, Cajamarca 06002, Perú

¶Centro de Investigaciones e Innovaciones de la Agroindustria Peruana. Amazonas 1061, Perú

Abstract—Worldwide, the monitoring of pests and diseases plays a fundamental role in the agricultural sustainability; making necessary the development of new tools for early pest detection. In this sense, we present a software application for detecting damage in tobacco (*Nicotiana tabacum* L.) leaves caused by the fungus of blue mold (*Peronospora tabacina* Adam). This software application processes tobacco leaves images using a pattern recognition technique known as Artificial Neural Network. For the training and testing stages, a total of 40 images of tobacco leaves were used. The experimentation carried out shows that the developed model has accuracy higher than 97% and there is no significant difference with a visual analysis carried out by experts in tobacco crop.

Keywords—*Nicotiana tabacum* L.; *Peronospora tabacina* Adam, image processing; artificial neural networks

I. INTRODUCTION

Worldwide, there is considerable interest in working towards agricultural sustainability. Therefore, the pests and diseases that attack different crops are some of the enormous challenges that must be faced. For this reason, integrated pest control is usually one of the most used approaches, where the monitoring of pests and diseases plays a fundamental role.

Tobacco (*Nicotiana tabacum* L.) is an economically important crop worldwide, which is grown in different agricultural areas around the world [1]. According to the SIAP¹, in 2015, 7,936 ha were planted in Mexico, which 1,264 ha were reported as damaged.

Blue mold (*Peronospora tabacina* Adam) is considered the main phytosanitary problem of tobacco cultivation in many countries worldwide [2]. The humid and cold climates favor the development of the fungus, which can be germinated in a period from 2 to 4 hours. This fungus can infect the tobacco plant in all its phenological stages. Initially, several localized lesions appear in different areas of the infected leaves, as the disease progresses, these lesions gather forming necrotic zones of light brown color. Fig. 1 shows a leaf of tobacco damaged by blue mold in its initial stage.

In general, farmers and agriculture experts detect diseases of their crops through visual inspection, which is based on their experience. This type of work requires continuous monitoring



Fig. 1. Tobacco leaf damaged by blue mold.

of expert people, which is often prohibitively expensive for many farmers due to the vast extensions of the plots. In several countries, there is government support for the monitoring of crops in order to guarantee agrifood sustainability. In Mexico, there is economic and technical support from some government agencies to help farmers monitor some pests and diseases^{2,3}. However, this evaluation process is tedious, time-consuming and moreover very much subjective. For this reason, the detection of diseases based on the processing of images is a very relevant research topic [3]. This technique allows non-expert people, helped by a software application, automatically detect the signals of diseases that occur in the crops.

The process of plant diseases recognition using images consists of extracting the feature information of the diseased regions in the obtained images by the use of image processing techniques, and then achieve the identification of diseases by employing some pattern recognition technique such as Neural Networks (NNs) [4], [5], Support Vector Machine (SVM) [6], [7], Genetic Algorithms (GA) [8], among others.

In particular, artificial neural networks (ANNs) stand out as a popular machine learning technique due to their relative ease of use and understanding compared to statistical methods, as well as their excellent performance in various machine learning tasks. In recent years, several studies have shown that ANNs have unique abilities such as learning and generalization which facilitates a reliable diagnosis of plant diseases [9], see Table I; in general, ANNs have a higher degree of diagnostic than other machine learning techniques.

²Mexican Service for Agroalimentary Public Health, Safety and Quality (Servicio Nacional de Sanidad, Inocuidad y Calidad Agroalimentaria). Available: <https://www.gob.mx/senasica>

³Nayarit Plant Health Committee (Comité Estatal de Sanidad Vegetal de Nayarit). Available: <http://cesavenay.org.mx/>

¹Mexican Agrifood and Fisheries Information System (Servicio de Información Agroalimentaria y Pesquera). Available: <http://infosiap.siap.gob.mx>

TABLE I. STRATEGIES FOR THE RECOGNITION OF DISEASES IN PLANTS BASED ON IMAGE PROCESSING AND NEURAL NETWORKS TECHNIQUES

Description	Technique	Accuracy	Reference
It is introduced that an algorithm for extracting lesion area and application of Probabilistic Neural Network (PNN) to classify seedling diseases such as anthracnose and frog-eye spots on tobacco leaves.	PNN	88.59%	[10]
It is presented a method to recognize wheat and grape diseases which is based on the principal component analysis (PCA) and PNN techniques. PCA is used to reduce the dimensions of the feature data to reduce the number of neurons in the input layer and to increase the speed of PNN.	PCA and PNN	95%	[4]
The authors introduced an image processing method to detect diseases which attack the pomegranate crop. The method is based on the artificial neural network (ANN) technique.	ANN	91%	[11]
Sannakki et al. presented a model to classify diseases that attack the leaves of the grape, in specific, downy mildew and powdery mildew. The model was based on the ANN technique.	ANN	100%	[5]
A method for identifying fungal diseases that affect vegetable crops is presented, the approach is based on the ANN technique.	ANN	84.11%	[12]

In this paper, with the intention of promoting the early identification of pests and the levels of damage in tobacco crops, a mobile application it is introduced to achieve the detection of damage in tobacco leaves caused by blue mold. A classification model was trained to boost this application, which is based on the multilayer perceptron ANN technique.

II. MATERIALS AND METHODS

A. Tobacco Leaves

In Mexico, more than 80% of tobacco crop is grown in the state of Nayarit, mainly in its coastal municipalities; being *Santiago Ixcuintla* the municipality where more than 70% of Nayarita tobacco is sown [13]. Fig. 2 shows the Nayarit state in solid green color, the leading tobacco region is highlighted in yellow tones, and the municipality of *Santiago Ixcuintla* is in intense solid yellow.

The tobacco leaves used in this study came from tobacco plantations located in the municipality of *Santiago Ixcuintla*; There, we identified healthy plants and plants with signs of damage by blue mold.

B. Protocol for the Acquisition of Images

The acquisition of the images was made in the following way: activating the focus option of the mobile device, taking a photograph at around 30 cm distance, making sure that there is proper solar illumination and absence of shadows and non-uniformities in light distribution over the leaf.

C. Samples

The data used in this research work consists of 40 images of tobacco leaves with different levels of damage. The photos

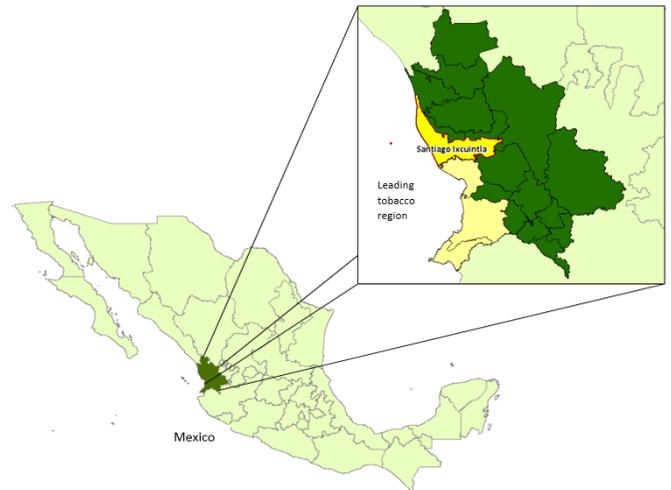


Fig. 2. Santiago Ixcuintla (highlighted in intense yellow color), municipality of the state of Nayarit (solid green color), the leader in the production of tobacco in Mexico.

were taken in different tobacco plantations located in Nayarit, Mexico.

D. Proposed Software System

This section outlines the overall scheme of the developed software application and the underlying technique. Fig. 3 shows the architecture of the proposed software system, which follows the client-server principle. On the client side, there are two components (1) a mobile application for the user interaction and (2) a local database. On the server side, there are three components, (3) a web service to address the requests of clients, (4) a database, and (5) a classification model based on the ANN technique. Next, each of the components is described.

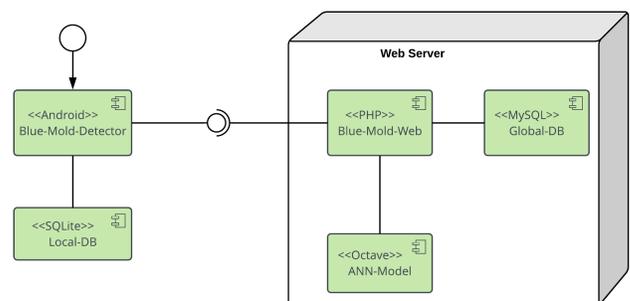


Fig. 3. The architecture of the proposed software application.

1) *Blue-Mold-Detector*: This software component is a mobile application based on the Android system. The user interacts with this component to take photographs of tobacco leaves; the photographs are sent to the central server for analysis; finally, the results are displayed on the mobile device.

2) *Local-DB*: In the fieldwork, there is often no Internet signal, for this reason, it is necessary to store the photographs taken on the mobile device and then when there is phone signal send them for analysis. The Local-DB component, based on SQLite software, is responsible for doing these tasks.

3) *Blue-Mold-Web*: This component is responsible for addressing the client's requests to analyze the images in search of blue mold. For this, this component interacts with the Global-DB and ANN-Model components. This component is a web service implemented in PHP.

4) *GlobalDB*: This component has the task of storing the images taken with the mobile application so that they can be processed later. The images are stored in a database, based on MySQL software.

5) *ANN classification model*: This module is responsible for analyzing the images sent by the Mold-Blue-Web component. This component is the classification model which is based on a multilayer perceptron neural network. The following section describes the method used to create the classification model based on the ANN technique.

E. Proposed Method for Developing the ANN Model

The following steps were followed to develop the proposed classification model: preprocessing, feature extraction, creation, and validation.

1) *Preprocessing*: The images were enhanced by a Gaussian filter; the function is shown in (1) was used for this purpose.

$$g(x, y) = \frac{1}{2\pi\sigma} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \tag{1}$$

where

g = Filtered image

x, y = Position of pixel

σ = Standard deviation of Gaussian filter.

2) *Feature extraction*: Similar to that presented by Castro et al. [14], a visual evaluation scale was established and three levels of tobacco leaf damage caused by blue mold are proposed. In Table II, the three levels of damage are described, from a healthy stage, through initial damage, to advanced damage showing large necrotic areas.

TABLE II. VISUAL SCALE FOR EACH LEVEL OF DAMAGE IN TOBACCO LEAVES CAUSED BY BLUE MOLD

Level	Stage	Visual aspect
1	Healthy	No visible stain.
2	Initial	Few yellow spots without defined borders.
3	Advance	The spots turn dark brown.

Subsequently, an application was developed in GNU Octave⁴ to extract characteristics of each level of damage. This application allows it to select manually rectangular regions of interest (ROIs) and associate the value of the RGB combinations of each pixel to a level of damage. In Fig. 4, the three histograms of blue mold damage levels are represented.

⁴GNU Octave. A Scientific Programming Language. Available: <https://www.gnu.org/software/octave/>

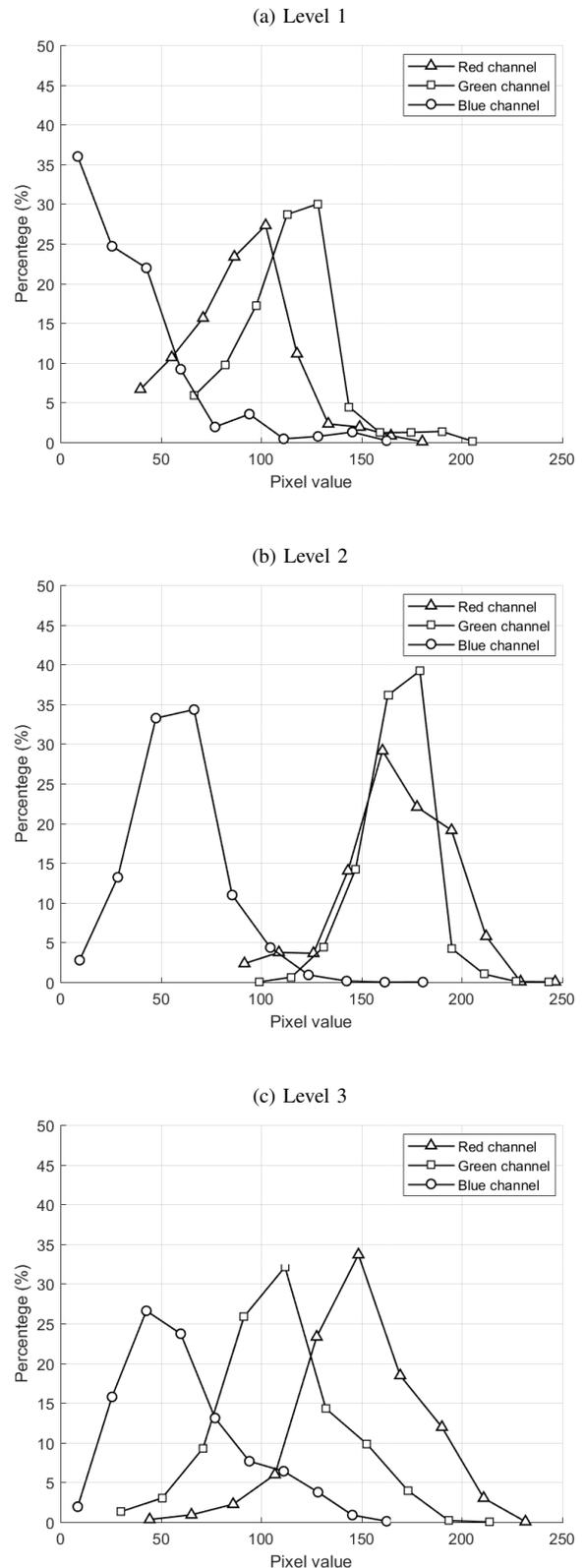


Fig. 4. Histograms of the different levels of damage.

3) *Creation and validation of the classification model*: In order to create and validate a classification model based on the ANN technique, the dataset was divided as follows: 50% for the creation and 50% for validation.

- **Creation.** In this step, we used the methodology suggested by Castro et al. [15], which were proposed for hyperspectral images; and we modified it to use tri-band images (RGB) and create a multilayer perceptron neural network, see Fig. 5.

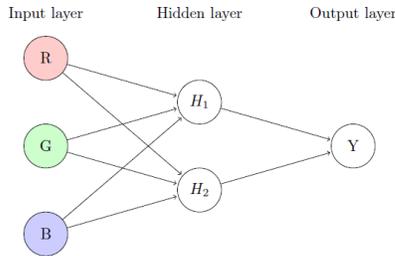


Fig. 5. The multilayer perceptron neural network architecture proposed.

In the proposed model, the number of neurons in the hidden layer was determined by the pyramidal principle given in (2). Three layers with three neurons in the input layer (one for each RGB band) and one neuron in the output layer with three possible values, one for each class (healthy, initial and advanced).

$$H_n = \frac{X_n + Y_n}{2} \quad (2)$$

where

H_n = hidden neurons

X_n = input neurons

Y_n = output neurons.

Once the algorithm based on the ANN technique was developed, 50% of the dataset was used as follows: training (70%), validation (15%), and tests (15%).

- **Validation.** In the same way, as in the previous step, the other 50% of the data set was used to evaluate the accuracy of the classification model created, the results are projected in a confusion matrix.

4) *Testing:* All the leaves were evaluated visually by a panel of four trained judges who determined the total area damaged. These values were compared with those obtained through the computational model, using descriptive statistics techniques.

III. RESULTS

A. Android APP

A prototype of a mobile application based on the Android system was developed, it offers an option to analyze the photographs captured on the secure digital card of the mobile device, see Fig. 6(a). Once the image is analyzed the results are shown in Fig. 6(b), it includes detailed information about the analysis, the date and time of the capture, the device that took the photograph, the GPS location of the leaf that was sent to be analyzed, and a graph of the levels of damage due to blue mold.

B. Obtaining Information by Level of Damage

The ROI selector system described in Section II-E2 was used to obtain the information by level of damage of tobacco

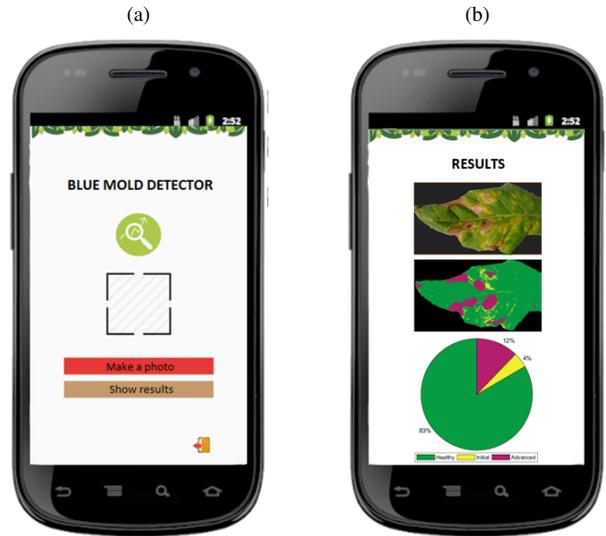


Fig. 6. The mobile software application proposed. (a) Main screen to take a photo of tobacco leaf and analyze it for damage caused by blue mold. (b) It shows the results of the analysis of a tobacco leaf.

leaves. From the obtained RGB combinations, we uniformly select 30,000 data (pixels) per damage level. As a result of the training process, an accuracy of 97.2% was obtained, see the confusion matrix shown in Table III.

TABLE III. CONFUSION MATRIX RESULTING FROM THE TRAINING PROCESS

	Healthy	Initial	Advanced	Accuracy
Healthy	33.2%	1.8%	0.8%	92.7%
Initial	0.1%	31.5%	0.0%	99.6%
Advanced	0.0%	0.0%	32.5%	100%
				97.2%

An example of the use of the developed classification model is shown in Fig. 7; (a) shows a classified leaf with different levels of damage, and (b) shows a graph which represents the percentage of damage for each level.

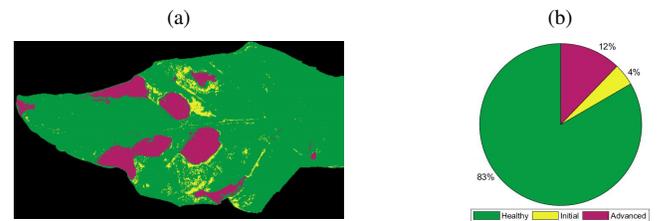


Fig. 7. An example of the classification model developed. (a) A leaf with different levels of damage. (b) A corresponding graph with the damage percentages for each level.

In order to assess the quality of the results obtained by the classifying model, we proceeded to compare the results against the reported in the literature. Table I shows the accuracy achieved by other classifiers based on neural networks techniques.

The only previous study that we found recorded in the literature regarding the use of automatic learning techniques applied to the detection of diseases and pests in tobacco crops

was reported by Guru et al. [10] in 2011. However, Guru et al. worked with tobacco seedlings and the damage caused by anthracnose. They proposed an algorithm to recognize the damaged areas, which was based on the technique of the probabilistic neural network (PNN). The developed classification model was able to detect and classify damages in leaves of tobacco seedlings with an accuracy of 88.58%. Therefore, the results we obtained in the classification of damage levels in tobacco leaves caused by blue mold using the ANN technique are very competitive (accuracy = 97.2%).

Wang et al. [4] used the PNN technique to detect diseases in wheat and grapes, obtaining accuracies of 95% and 94.29% respectively. Concerning the sample size, they used an average of 25 images for each disease. The results that we report in this paper using a sample of 40 images are superior to those reported by Wang et al.

Regarding the specific use of the ANN technique in the classification of diseases in crops, we find in the literature the works of Kulkarni and Patil [11] (pomegranate), Sannakki et al. [5] (grape), and Pujari et al. [12] (beans, soybean, sunflower, and tomato). In summary, our results are very competitive compared to those reported in the literature and that only the results reported by Sannakki et al. are superior, see Table I.

Analyzing the work reported by Sannakki et al. [5], we found that the addressed problem could be considered more accessible to tackle compared to the problem proposed in this paper. Sannakki et al. addressed the problem of identifying between two types of diseases that attack the leaves of grapes (downy mildew and powdery mildew); instead, we address the problem of determining levels of damage caused by the blue mold in tobacco leaves. A detail that caught our attention is that Sannakki et al. mentioned that they used 94% of their data for training and only 6% for validation, which could have positively influenced their results.

C. System Testing

First, four judges specialized in tobacco pests carried out a visual analysis (VA). They made a visual examination of the tobacco leaf samples using the scale showed in Table II.

Due to the differences in the results of the applied methods (ANN and VA), it is necessary to determine if such differences are significant. For this purpose, a nonparametric statistical analysis was performed using the Kolmogorov-Smirnov test, which we calculated using (3).

$$D = \max |VA_i - ANN_i| = 0.2023 \quad (3)$$

For $\alpha = 0.05$, we obtain a $D_\alpha = 0.2101$, since $D < D_\alpha$ it is concluded that there is no significant difference in the results.

IV. CONCLUSIONS

In this work, a mobile application was presented for the detection of damage levels in tobacco leaves due to blue mold. Internally, this mobile application has a damage classification model based on the ANN technique. The experimentation carried out, showed that this model has an accuracy superior to 97% and there is no significant difference with a visual

analysis made by experts in pests and diseases of the tobacco. As future work, we will extend the use of the proposed APP to identify other diseases and pests in tobacco crops.

ACKNOWLEDGMENT

The authors would like to thank the technical staff of the *CONTMIPEBIO S.P.R. de R.L.* company for all the help given, both professional and logistical when it comes to collecting the samples.

REFERENCES

- [1] S. Sukanya and O. Spring, "Influence of temperature and ultra-violet light on viability and infectivity of peronospora tabacina sporangia," *Crop Protection*, vol. 51, pp. 14 – 18, 2013.
- [2] O. Spring, T. Hammer, R. Zipper, and N. Billenkamp, "Population dynamics in tobacco blue mold incidences as a consequence of pathogen control and virulence performance of peronospora tabacina phenotypes," *Crop Protection*, vol. 45, pp. 76 – 82, 2013.
- [3] V. Singh and A. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," *Information Processing in Agriculture*, vol. 4, no. 1, pp. 41 – 49, 2017.
- [4] H. Wang, G. Li, Z. Ma, and X. Li, "Image recognition of plant diseases based on principal component analysis and neural networks," in *Natural Computation, 2012 Eighth International Conference on*. IEEE, 2012, pp. 246–251.
- [5] S. S. Sannakki, V. S. Rajpurohit, V. B. Nargund, and P. Kulkarni, "Diagnosis and classification of grape leaf diseases using neural networks," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies*. IEEE, 2013, pp. 1–5.
- [6] A. Camargo and J. S. Smith, "Image pattern classification for the identification of disease causing agents in plants," *Computers and Electronics in Agriculture*, vol. 66, no. 2, pp. 121–125, 2009.
- [7] T. Rumpf, A.-K. Mahlein, U. Steiner, E.-C. Oerke, H.-W. Dehne, and L. Plümer, "Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance," *Computers and Electronics in Agriculture*, vol. 74, no. 1, pp. 91–99, 2010.
- [8] M. Arshad, S. Ullah, K. Khurshid, and A. Ali, "Estimation of leaf water content from mid-and thermal-infrared spectra by coupling genetic algorithm and partial least squares regression," *Journal of Applied Remote Sensing*, vol. 12, no. 2, p. 022203, 2018.
- [9] K. Golhani, S. K. Balasundram, G. Vadamalai, and B. Pradhan, "A review of neural networks in plant disease detection using hyperspectral data," *Information Processing in Agriculture*, 2018.
- [10] D. S. Guru, P. B. Mallikarjuna, and S. Manjunath, "Segmentation and classification of tobacco seedling diseases," in *Proceedings of the Fourth Annual ACM Bangalore Conference*. ACM, 2011, p. 32.
- [11] A. H. Kulkarni and A. Patil, "Applying image processing technique to detect plant diseases," *International Journal of Modern Engineering Research*, vol. 2, no. 5, pp. 3661–3664, 2012.
- [12] J. D. Pujari, R. Yakkundimath, and A. S. Byadgi, "Image processing based detection of fungal diseases in plants," *Procedia Computer Science*, vol. 46, pp. 1802–1808, 2015.
- [13] SIAP, "Atlas agroalimentario 2016," Mexican Agrifood and Fisheries Information System (Servicio de Información Agroalimentaria y Pesquera), 2016. [Online]. Available: https://nube.siap.gob.mx/gobmx_publicaciones_siap/pag/2016/Atlas-Agroalimentario-2016
- [14] W. Castro, J. Oblitas, J. Maicelo, and H. Avila-George, "Evaluation of expert systems techniques for classifying different stages of coffee rust infection in hyperspectral images," *International Journal of Computational Intelligence Systems*, vol. 11, no. 1, pp. 86–100, 2018.
- [15] W. Castro, J. Oblitas, R. Santa-Cruz, and H. Avila-George, "Multilayer perceptron architecture optimization using parallel computing techniques," *PLoS ONE*, vol. 12, no. 12, p. e0189369, 2017.

Complex Shear Modulus Estimation using Integration of LMS/AHI Algorithm

Quang–Hai Luong, Manh–Cuong Nguyen
Le Quy Don Technical University
Hanoi, Vietnam

TonThat–Long
School of Electrical Engineering University of Engineering and Technology
International University
VNU-HoChiMinh City
HoChiMinh City, Vietnam

Duc–Tan Tran
University of Engineering and Technology
VNU, Hanoi
Hanoi, Vietnam

Abstract—Elasticity and viscosity of tissues are two important parameters that can be used to investigate the structure of tissues, especially detecting tumors. By using a force excitation, the shear wave speed is acquired to extract its amplitude and phase. This information is then used directly or indirectly to compute the Complex Shear Modulus (CSM consists of elasticity and viscosity). Among these methods, Algebraic Helmholtz Inversion (AHI) algorithm can be combined with the Finite Difference Time Domain (FDTD) model to estimate CSM effectively. However, this algorithm is strongly affected by measured noise while acquiring the particle velocity. Thus, we proposed a LMS/AHI algorithm which can estimate correctly CSM. A simulation scenario is built to confirm the performance of the proposed LMS/AHI algorithm with average error of 3.14%.

Keywords—Shear wave; elasticity; viscosity; CSM estimation; least mean square; Algebraic Helmholtz Inversion

I. INTRODUCTION

Elasticity and viscosity of tissues are two important factors that can be exploited to detect tumors [1]. Many research work focused on elasticity [2]–[7] where ultrasonic Shear Wave Elasticity Imaging (SWEI) offers significant advantages over the other techniques in terms of reproducibility, quantification, elasticity contrast, and automatic shear wave generation. However, for deeply understanding about the tissue, various work have been developed to estimate both the elasticity and viscosity, which are briefly surveyed next. Breast needle biopsy is well-known in ultrasonic. In order to generate the shear wave, previous work applied force whose frequencies are low as 0.1 Hz and high as 10 kHz. Recently, they have used the excitation in the range of 50–250 Hz for simplification of the measurement. In this paper, only single frequency of 150 Hz is needed for the excitation. By applying the force at different spatial locations, the structure of the tissue can be investigated.

The relationship between the speed and absorption of shear wave to the corresponding CSM can be modeled by simple equations in [8]. If transient forces [9] are considered, the reflections are minimized. However, the affection of noise is worse than using harmonic forces [10]. If the harmonic needle vibration is used, compared with other force excitation techniques, larger amplitudes of shear wave can be obtained. Thus, a harmonic needle vibration is considered for excitation in this work. In 2004, Chen *et al.* exploited the relationship between the propagation speed and the vibrating frequency to build the shear wave speed dispersion, and then estimate the CSM [11]. In 2007, Zheng *et al.* used a linear Kalman filter for

CSM estimation over a frequency bandwidth [10]. The noise is reduced by this filter. Recently, some extended methods have been introduced in [10]–[13] where the authors needed to use multiple datasets of different vibration frequencies. In this paper, only a single-frequency excitation is needed, but still, the acquisition time is improved.

In order to detect tumor (if any) in the tissue, Tran *et al.* [14] used the maximum likelihood ensemble filter for 1D heterogeneous tissue. However, the propagation model using wave equation in [12], [14], [15] is very simple, and it can not represent the actual propagation in heterogeneous tissue. In 2015, Qiuang *et al.* [16] proposed a method which uses Finite Element Method (FEM) to model the shear wave propagation in transversely isotropic, viscoelastic and incompressible media. However, the complexity of FEM is high. FDTD is more effective method with lower complexity than FEM. In [17], Orescanin *et al.* exploited FDTD model then used AHI algorithm to estimate CSM. However, there is a lack of deep investigation of noise in this work because AHI is strongly affected by noise.

In this paper, we introduce an integration of AHI and Least Mean Square (LMS) algorithms to estimate CSM. A shear wave generator at a single frequency of 100 Hz is excited at the origin (0,0) by the vibrating needle. A linear array transducer is used to measure the particle velocity of shear wave at 120 spatial locations. At each point, the CSM from the noisy particle velocity of shear wave is then estimated by applying a specific LMS/AHI. Using the LMS/AHI can drastically reduce the complexity as compared to previous techniques. Finally, a scenario with a tumor and noise environment is studied to evaluate quality of the estimated CSM.

II. METHOD

A. Shear Wave Propagation

Generation of shear waves and measurements of the particle velocity are shown in Fig. 1. In this system, a needle is vibrated at a single frequency along the Z-axis, the share wave is then propagated in X-Y plane. After that, the particle velocity is acquired by using a Doppler ultrasound device [12].

In some previous work ([12], [14], [15]), the wave (1) is used to compute the particle velocity $v(r, t)$ at a spatial

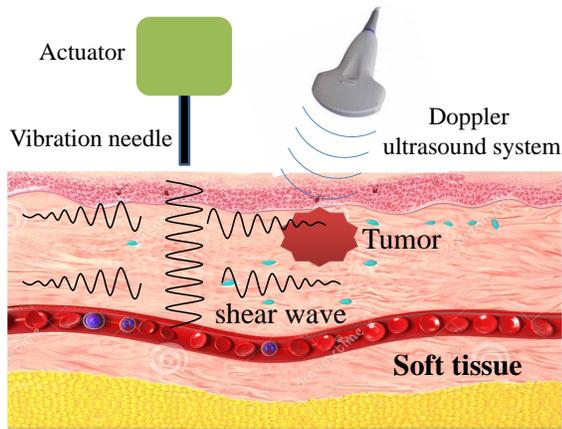


Fig. 1. Generation and measurement of shear wave.

location r and time t .

$$v(r, t) = \frac{1}{\sqrt{r - r_0}} A e^{-\alpha(r - r_0)} \cos[\omega t - k_s(r - r_0) - \phi], \quad (1)$$

where A is the vibration's amplitude of the needle, r_0 is the needle's spatial location, ϕ is the initial phase, α and k_s are attenuation coefficient and wave number at spatial location r respectively.

The formula (1) has the advantage of simplification. However, it cannot reflect the propagation of the shear wave in the real tissue, especially in a heterogeneous one. Thus, FDTD method is used, together with the assumption of cylindrical shear wave propagation along the radial axis and ignoring absorption of medium. Consequently, the particle velocity vector v_z on a direction of the wave propagation x in Cartesian coordinate relates to the stress tensor σ_{zx} , which can be described by the following (2) and (3) (from [17]):

$$\rho \partial_t v_z = \partial_x \sigma_{zx}, \quad (2)$$

$$\partial_t \sigma_{zx} = (\mu + \eta \partial_t) \partial_x v_z, \quad (3)$$

where ∂_t represents a partial derivative operator $\partial/\partial t$ applied to values to the right of the symbol, ∂_x represents a partial derivative operator $\partial/\partial x$ applied to values to the right of the symbol, ρ is density of the tissue, μ and η are the elasticity and viscosity of the tissue respectively.

Kelvin-Voigt model is applied to represent the CSM $G(x, \omega)$, which depends on the angle frequency of the vibration ω as follows

$$G(x, \omega) = \mu(x) - i\omega\eta(x). \quad (4)$$

where μ is the elasticity and η is the viscosity that need to be estimated. To discretize (2) and (3), the following notations will be used:

$$v_z(x, t) = v_z(i\Delta x, n\Delta t) = v_z^n |_i, \quad (5)$$

$$\sigma_{zx}(x, t) = \sigma_{zx}(i\Delta x, n\Delta t) = \sigma_{zx}^n |_i, \quad (6)$$

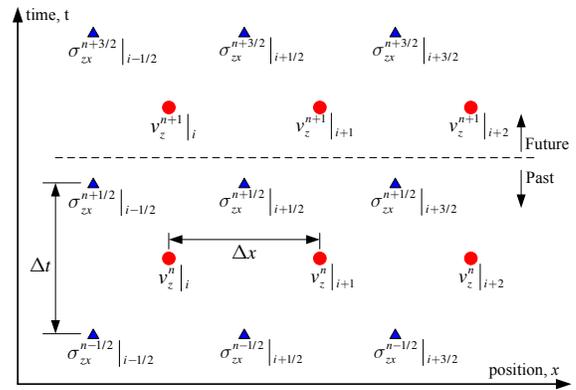


Fig. 2. Illustration of the stress tensor σ and the particle velocity vector v_z nodes in time and space.

where Δx is the distance between continuous spatial locations, Δt is the sampling period, the index i is the spatial step, and the index n is the temporal step, as shown in Fig. 2.

By using FDTD method, (2) and (3) are described as follows:

$$v_z^{n+1} |_i = v_z^n |_i + \frac{\Delta t}{\rho \Delta x} \left(\sigma_{zx}^{n+\frac{1}{2}} |_{i+\frac{1}{2}} - \sigma_{zx}^{n+\frac{1}{2}} |_{i-\frac{1}{2}} \right), \quad (7)$$

$$\begin{aligned} \sigma_{zx}^{n+\frac{1}{2}} |_{i+\frac{1}{2}} &= \sigma_{zx}^{n-\frac{1}{2}} |_{i+\frac{1}{2}} + \frac{\mu \Delta t}{\Delta x} (v_z^{n+1} |_{i+1} - v_z^{n+1} |_i) \\ &\quad + \frac{\eta}{\Delta x} (v_z^{n+1} |_{i+1} - v_z^{n+1} |_i) \\ &\quad - \frac{\eta}{\Delta x} (v_z^n |_{i+1} - v_z^n |_i), \end{aligned} \quad (8)$$

B. Signal Enhancement using Least Mean Square Algorithm

Adaptive filters attracts a great attention due to its property of self adjusting their coefficients [18]. For the signal enhancement, the output signal is obtained from a noisy input signal and an adaptation process. The filter coefficients are adjusted in order to minimize a desired cost function. There are a lot of filter structures and adaptive algorithms that have been developed in recent decades [19]. In this paper, we design a transversal adaptive filter to reduce the noise from the noisy particle velocity which is acquired from the Doppler ultrasound system, as shown in Fig. 3.

In Fig. 3, a particle velocity signal, represented as $d(n)$, is transmitted into the tissue which is affected by noises, represented as $v(n)$. Together, they form a noisy signal $v_z(n)$ which is described by

$$v_z(n) = d(n) + v(n). \quad (9)$$

This noisy signal $v_z(n)$ is applied as an input to the adaptive filter to extract the estimate of the desired signal with minimum error using various adaptive methods such as LMS, Normalised Least Mean Square (NLMS), Root Mean Square (RMS) algorithms, etc. When the estimate of noise equals or approximates the $v[n]$ ($y(n) = v(n)$), the error

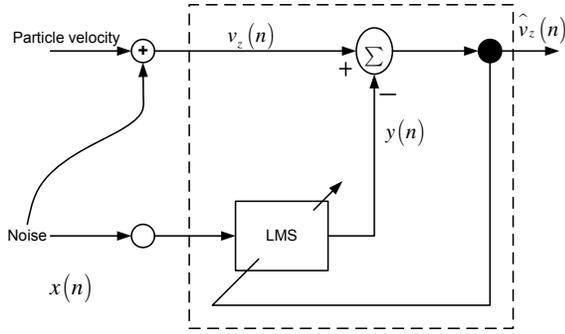


Fig. 3. Using LMS filter to enhance the particle velocity.

signal is approximately the filtered speech signal $d(n)$ because $e(n) = d(n) + v(n) - v(n) = d(n)$. The output of the adaptive filter or filtered signal $\hat{v}_z(n) = e(n)$.

The input signal is sampled and it forms a vector containing N samples.

$$v_z(n) = [v_z(0) \ v_z(1) \ \dots \ v_z(N-1)], \quad (10)$$

The corresponding estimated value of the desired signal is $\hat{v}_z(n)$. The coefficients of the filter are represented as

$$w(n) = [w(0) \ w(1) \ \dots \ w(L)]. \quad (11)$$

where L represents the order of the filter. The filter coefficients, alternatively called as weights $w(n)$, are adjusted every time in such a way that the Mean Square Error (MSE) is minimized. There are many well-known adaptive algorithms that can accomplish this weight adjustment. Among them, the LMS algorithm has a simple filter weight update mechanism, which has a fast rate of convergence if an optimal step size is used.

The LMS algorithm can be summarized in *Algorithm 1* below:

Algorithm 1: LMS Algorithm for Shear Wave Acquisition

- Step 1:** Initialize the step size μ , filter order L , and noise variance.
- Step 2:** Initialize the filter coefficients $w(n) = 0$.
- Step 3:** For $n = 0, 1, 2, \dots$
 - 3.1. Compute the filter output $y(n) = w(n) * x(n)$
 - 3.2. Compute the error in estimation $e(n) = v_z(n) - y(n)$
 - 3.3. Compute the updated tap-weight $w(n+1) = w(n) + \mu e(n)x(n)$
 - 3.4. Compute the denoised signal by assigning $\hat{v}_z(n) = e(n)$
 - 3.5. Iterate till end of the signal

end

C. Direct Inversion using Least Mean Square/Algebraic Helmholtz Inversion Algorithm to Estimate the CSM

After reducing the noise from acquired particle velocity, the AHI algorithm [20] is used to compute the CSM. For a small volume, it is assumed that the viscoelastic property of

the tissue is isotropic. Thus, there is negligible compression applied to the tissue by the needle, as a result, the particle velocity vector v_z can be described by the Navier wave equation in a homogeneous solid. We combine (2) and (3) to obtain

$$\rho \frac{\partial^2 v_z}{\partial t^2} = G'(x, t) \nabla^2 v_z, \quad (12)$$

where $G'(x, t)$ is the CSM in time domain and $\nabla^2 v_z$ is Laplace operator of v_z which is defined as $\nabla^2 v_z = \partial^2 v_z / \partial x^2$.

AHI algorithm is applied to solve (12), which then becomes the Helmholtz equation

$$\left(\frac{G(x, \omega)}{\rho} \nabla^2 + \omega^2 \right) V_z(x, \omega) |_{\omega=\omega_0} = 0, \quad (13)$$

where $G(x, \omega)$ is the CSM in frequency domain and defined in (4), $V_z(x, \omega)$ is the temporal Fourier transform of the particle velocity $v_z(x, t)$, $V_z(x, \omega) = F_t \{v_z(x, t)\}$, and ω_0 is the angular frequency $\omega_0 = 2\pi f_0$. From (13), it can be seen that the CSM can be estimated directly as

$$\mu(x) = \Re \left\{ \frac{-\rho \omega_0^2 V_z(x, \omega_0)}{\nabla^2 V_z(x, \omega_0)} \right\}, \quad (14)$$

$$\eta(x) = \Im \left\{ \frac{-\rho \omega_0 V_z(x, \omega_0)}{\nabla^2 V_z(x, \omega_0)} \right\},$$

where $V_z(x, \omega_0)$ is computed by using Fourier transform at the specific angular frequency ω_0 ; $\nabla^2 V_z(x, \omega_0)$ is computed by using the function Discrete Laplacian (The MathWorks) $del2(V_z(x, \omega_0))$ which returns a discrete approximation of Laplaces differential operator applied to $V_z(x, \omega_0)$.

The proposed LMS/AHI for CSM estimation is summarized in *Algorithm 2*.

Algorithm 2: LMS/AHI Algorithm for CSM Estimation

- Step 1:** Set up the simulation scenario.
- Step 2:** Select the excitation frequency $f_0=150\text{Hz}$.
- Step 3:** Generate shear waves by vibrating the needle.
- Step 4:** Acquire the noisy particle velocity at 120 spatial locations.
- Step 5:** Estimate the noise variance from the noisy signal corrupted by the additive white noise.
- Step 6:** Reduce the noise from noisy particle velocity using LMS filter as shown in *Algorithm 1*.
- Step 7:** Discard the transient parts of the filtered signal and keep the steady-state of the filtered particle velocity.
- Step 8:** Compute the temporal Fourier transform of the filtered signal.
- Step 9:** Estimate each CSM in spatial locations using (14).
- Step 10:** Evaluate the estimation performance.

end

III. RESULTS AND DISCUSSIONS

In order to verify the proposed method LMS/AHI, a simulation scenario is built where 1D heterogeneous tissue is 12 mm in length and the distance between two continuous spatial locations is 0.1 mm. At each spatial location, we consider 20000 samples. The vibrating frequency is chosen as 150 Hz; the amplitude of the vibration is 2 mm. The elasticity and viscosity of the tissue are shown in Table I.

TABLE I. SIMULATION SCENARIO

Spatial locations	Elasticity [Pa]	Viscosity [Pa.s]
1–29	650	0.1
30–50	900	0.35
51–120	650	0.1

It can be seen that there is an inclusion (from the 30th to 50th spatial locations) in this tissue.

Fig. 4 and 5 present the particle velocities in term of time at the 15th and 60th points, respectively. They are sinusoidal functions of time in two cases: noisy and filtered signals. The filtered signal can be divided into two parts: transient part which is still affected by noise and the steady-state part where the noise was filtered out completely. It is obviously that the amplitude at the 15th spatial location is larger than that of the 60th one. However, the power of the additive noise is the same for all $L = 120$ spatial locations. It is noted that the noise would strongly affect the CSM estimation due to the limitation of AHI methodology.

Fig. 6 shows the particle velocity vs. spatial locations where the undeniable role of measured noise is also illustrated. Furthermore, as shown in this figure, the power of the additive noise is the same for every spatial location.

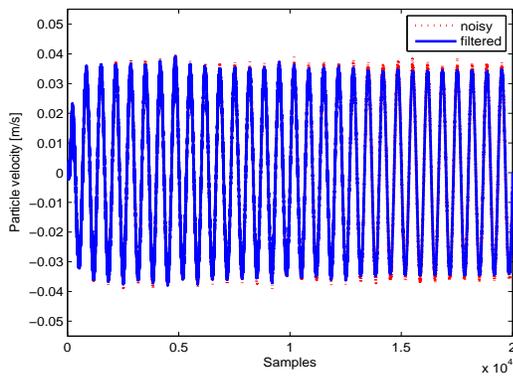


Fig. 4. The particle velocity in time at the 15th point.

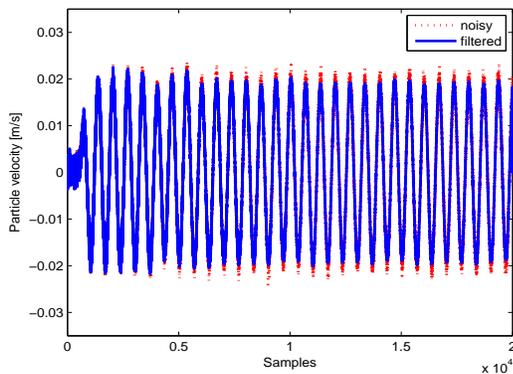


Fig. 5. The particle velocity in time at the 60th point.

Without noise filtering, CSM can not be estimated. Thus, the following three cases, which are shown in Fig 7 and 8, are concerned: ideal estimation, LMS/AHI without removing

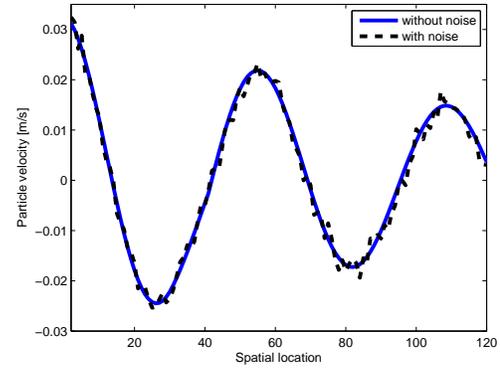


Fig. 6. The particle velocity vs. spatial locations.

the transient parts of the filtered signal, and the proposed LMS/AHI (i.e removing the transient part and keep the steady-state of the filter signal). Without removing the transient part, the estimation can not trace well the ideal ones. The reason is that the noise is still existed in this part and it affects both calculations of $v_z(x, \omega_0)$ and $\nabla^2 v_z(x, \omega_0)$. By cutting the transient parts of the filtered signal and keep the steady-state of the filtered particle velocity, the calculations of both $v_z(x, \omega_0)$ and $\nabla^2 v_z(x, \omega_0)$ are improved. Hence, the elasticity and viscosity can trace well the ideal ones.

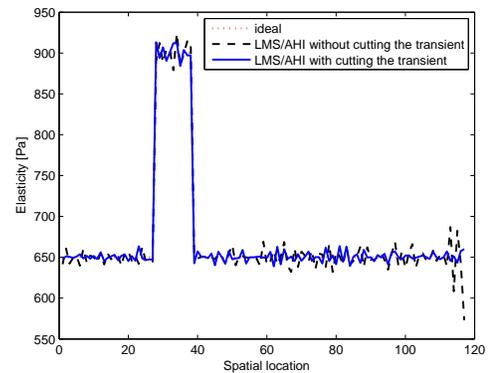


Fig. 7. The estimated elasticity.

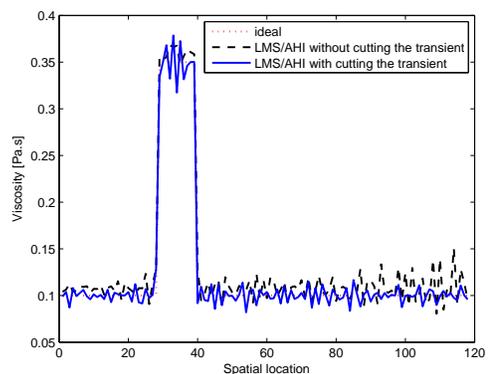


Fig. 8. The estimated viscosity.

To quantify the efficiency of the proposed LMS/AHI algorithm, the error between the ideal CSM (μ, η) and the estimated CSM ($\hat{\mu}, \hat{\eta}$) on different ranges of the tissue is computed. The normalized error can be defined as:

$$\begin{aligned}\epsilon_{\mu} &= \frac{1}{L} \sum_{i=1}^L \frac{|\mu_i - \hat{\mu}_i|}{\mu_i} \\ \epsilon_{\eta} &= \frac{1}{L} \sum_{i=1}^L \frac{|\eta_i - \hat{\eta}_i|}{\eta_i}\end{aligned}\quad (15)$$

The results are shown in Table II. It can be seen that the quality of estimation is significantly improved by cutting the transient part of the filtered particle velocity. The transient part, which is still affected by noise, would affect the CSM estimation. The efficiency of the proposed LMS/AHI algorithm is confirmed by this error performance. As indicated in Table II, with LMS/AHI algorithm after cutting the transient part, we can achieve the average error of both elasticity and viscosity as 3.14%.

TABLE II. ERROR OF THE CSM ESTIMATION

Spatial locations	Error for elasticity [%]	Error for viscosity [%]
LMS/AHI without cutting the transient	1.27	10.39
LMS/AHI with cutting the transient	0.64	5.64

The results in this study are compared with some previous results. In [14], [15] and [12], authors used the MLEF (Maximum Likelihood Ensemble Filter) to estimate CSM. The error of estimation in [14], [15] is less than 10% for both elasticity and viscosity while in [12], the error of elasticity estimation is less than 2% and the error of viscosity estimation is within 5%. However, in the studies [14], [15] and [12], the particle velocity of shear wave is modelled following the basic wave propagation equation which is only suitable for a homogeneous medium. In this paper, we have applied FDTD method to model the particle velocity of shear wave and this is correct with a heterogeneous medium (i.e. tissues). In addition, an adaptive filter (LMS) has been added for denoising the measured particle velocity of shear wave before estimating CSM using AHI algorithm. This is an advantage of this paper comparing with [17] which applied only AHI algorithm to estimate CSM.

IV. CONCLUSION

A method for 1D estimation of CSM in tissues have been proposed successfully in this paper. The proposed method has used LMS/AHI to estimate CSM at each point in the medium with good estimation error of 3.14%. In the experiment, only a single vibration frequency (150 Hz) is used. The propagation of shear wave is modeled using FDTD method with good accuracy and low complexity compared to FEM. The noise in the particle velocity is filtered using a specific LMS filter. CSM is then estimated directly using AHI after removing the transient part of filtered signal. In the future work, we will improve the accuracy of the CSM estimation, especially for the viscosity of tissues. The target will be able to distinguish between objects and the medium when the CSM of the objects are not much greater than that of the medium (i.e. early tumor detection).

ACKNOWLEDGMENT

This work was supported by the Asia Research Center (ARC), Vietnam National University, code CA.17.6A.

REFERENCES

- [1] J. Bercoff, A. Cifre, C. Bacrie, J. Souquet, M. Tanter, J. Gennisson, T. Deffieux, M. Fink, V. Juhan, A. Colavolpe *et al.*, "ShearWave Elastography A new real time imaging mode for assessing quantitatively soft tissue viscoelasticity," in *Ultrasonics Symposium, 2008*. IEEE, 2008, pp. 321–324.
- [2] A. P. Sarvazyan, O. V. Rudenko, S. D. Swanson, J. B. Fowlkes, and S. Y. Emelianov, "Shear wave elasticity imaging: a new ultrasonic technology of medical diagnostics," *Ultrasound in medicine & biology*, vol. 24, no. 9, pp. 1419–1435, 1998.
- [3] J.-L. Gennisson, T. Deffieux, M. Fink, and M. Tanter, "Ultrasound elastography: principles and techniques," *Diagnostic and interventional imaging*, vol. 94, no. 5, pp. 487–495, 2013.
- [4] G. Ferraioli, P. Parekh, A. B. Levitov, and C. Filice, "Shear wave elastography for evaluation of liver fibrosis," *Journal of Ultrasound in Medicine*, vol. 33, no. 2, pp. 197–203, 2014.
- [5] Y. Kobayashi, M. Tsukune, T. Miyashita, and M. G. Fujie, "Simple empirical model for identifying rheological properties of soft biological tissues," *Physical Review*, vol. 95, no. 2, 2017.
- [6] S. Woo, S. Y. Kim, M. S. Lee, J. Y. Cho, and S. H. Kim, "Shear wave elastography assessment in the prostate: an intraobserver reproducibility study," *Clinical imaging*, vol. 39, no. 3, pp. 484–487, 2015.
- [7] W. Zhang and S. Holm, "Estimation of shear modulus in media with power law characteristics," *Ultrasonics*, vol. 64, pp. 170–176, 2016.
- [8] J. F. Greenleaf, M. Fatemi, and M. Insana, "Selected methods for imaging elastic properties of biological tissues," *Annual review of biomedical engineering*, vol. 5, no. 1, pp. 57–78, 2003.
- [9] L. Sandrin, B. Fourquet, J.-M. Hasquenoph, S. Yon, C. Fournier, F. Mal, C. Christidis, M. Ziol, B. Poulet, F. Kazemi *et al.*, "Transient elastography: a new noninvasive method for assessment of hepatic fibrosis," *Ultrasound in Medicine and Biology*, vol. 29, no. 12, pp. 1705–1713, 2003.
- [10] Y. Zheng, S. Chen, W. Tan, R. Kinnick, and J. Greenleaf, "Detection of tissue harmonic motion induced by ultrasonic radiation force using pulse-echo ultrasound and kalman filter," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 54, no. 2, pp. 290–300, 2007.
- [11] S. Chen, M. Fatemi, and J. F. Greenleaf, "Quantifying elasticity and viscosity from measurement of shear wave speed dispersion," *The Journal of the Acoustical Society of America*, vol. 115, no. 6, pp. 2781–2785, 2004.
- [12] M. Orescanin and M. F. Insana, "Model-based complex shear modulus reconstruction: A Bayesian approach," in *Ultrasonics Symposium*. IEEE, 2010, pp. 61–64.
- [13] Y. Wang and M. F. Insana, "Viscoelastic properties of rodent mammary tumors using ultrasonic shear-wave imaging," *Ultrasonic imaging*, vol. 35, no. 2, pp. 126–145, 2013.
- [14] T. Tran-Duc, Y. Wang, N. Linh-Trung, M. N. Do, and M. F. Insana, "Complex Shear Modulus Estimation Using Maximum Likelihood Ensemble Filters," in *4th International Conference on Biomedical Engineering in Vietnam*. Springer Berlin Heidelberg, 2013, pp. 313–316.
- [15] N. T. Hao, T. Thuy-Nga, V. Dinh-Long, T. Duc-Tan, and N. Linh-Trung, "2D Shear Wave Imaging Using Maximum Likelihood Ensemble Filter," in *International Conference on Green and Human Information Technology (ICGHIT)*, 2013, pp. 88–94.
- [16] B. Qiang, J. Brigham, S. Aristizabal, J. Greenleaf, X. Zhang, and M. Urban, "Modeling transversely isotropic, viscoelastic, incompressible tissue-like materials with application in ultrasound shear wave elastography," *Physics in medicine and biology*, vol. 3, no. 60, pp. 1289–1306, 2015.

- [17] M. Orescanin, Y. Wang, and M. F. Insana, "3d ftdt simulation of shear waves for evaluation of complex modulus imaging," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 58, no. 2, pp. 389–398, 2011.
- [18] S. Haykin, "Adaptive filter theory," 2008.
- [19] M. H. Hayes, *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009.
- [20] S. Papazoglou, U. Hamhaber, J. Braun, and I. Sack, "Algebraic helmholtz inversion in planar magnetic resonance elastography," *Physics in medicine and biology*, vol. 53, no. 12, p. 3147, 2008.

RASP-TMR: An Automatic and Fast Synthesizable Verilog Code Generator Tool for the Implementation and Evaluation of TMR Approach

Abdul Rafay Khatri, Ali Hayek, and Josef Börcsök
Department of Computer Architecture and System Programming,
University of Kassel, Kassel, Germany

Abstract—Triple Modular Redundancy (TMR) technique is one of the most well-known techniques for error masking and Single Event Effects (SEE) protection for the FPGA designs. These FPGA designs are mostly expressed in hardware description languages, such as Verilog and VHDL. The TMR technique involves triplication of the design module and adding the majority voter circuit for each output port. Building this triplication scheme is a non-trivial task and requires a lot of time and effort to alter the code of the design. In this paper, the RASP-TMR tool is developed and presented that has functionalities to take a synthesizable Verilog design file as an input, parse the design and triplicate it. The tool also generates a top-level module in which all three modules are instantiated and finally adds the proposed majority voter circuit. This tool, with its graphical user interface, is implemented in MATLAB. The tool is simple, fast and user-friendly. The tool generates the synthesizable design that facilitates the user to evaluate and verify the TMR design for FPGA-based systems. A simulation scenario is created using Xilinx ISE tools and ISim simulator. Different fault models are examined during simulations such as bit-flip and stuck at 1/0. The results using various benchmark designs demonstrate that the tool produces synthesizable code and the proposed majority voter logic perfectly masks the error/failure.

Keywords—Fault injection; fault tolerance; reliability; single event effects; triple modular redundancy; Verilog HDL

I. INTRODUCTION

The Field Programmable Gate Array (FPGA) has been a widely accepted solution in developing the embedded system during the last few decades. Owing to its remarkable features such as parallelism, reconfiguration, separation of functions, self-healing capabilities, and overall availability [1], the FPGA has become the core of many embedded applications. The major applications include aerospace, biomedical instrumentation, safety-critical systems, spacecraft, Internet of Things (IoT), and many others [2], [3]. However, FPGA-based devices are susceptible to Single Event Effects (SEE) caused by various sources such as α -particles, cosmic rays, atmospheric neutrons, and heavy-ion radiations. Since the capacity of FPGA chip has been growing by reducing the size of components integrated on the chip. This makes the device more prone to SEEs which provoke Single Event Upsets (SEU) in memory elements and Single Event Transients (SET) in combinational logic elements [4], [5].

Several SEE mitigation techniques have been proposed in the literature to avoid the effects of such errors in FPGA-based designs [6]. The reliability of the FPGA systems is im-

proved by various error mitigation schemes such as multiple-redundancy with voting, Triple Modular Redundancy (TMR), hardened memory cell level, and Error Detection And Correction (EDAC) coding. Among all SEU mitigation techniques, TMR has become the most common practice because of its straightforward implementation and achieved the reliable results [6], [7], [8], [9]. The TMR mitigation scheme uses three identical logic circuits for performing the same task in parallel with corresponding outputs obtained through majority voters. This technique is the simplest redundant technique conceptually and it was invented by Von Neumann [10] in the year 1956. There are certain merits and demerits of TMR technique:

- Merits:
 - Simple and straight forward approach.
 - Improves fault tolerance and reliability.
- Demerits:
 - Large area overhead, nearly 200%.
 - More power consumption.
 - More cost or uneconomical.

FPGA designs are written in Hardware Description Languages (HDL) which describes the designs in various abstraction style, for example, gate, data-flow and behavioural levels. For small designs, gate abstraction style is employed and testing & verification processes are directly and easily applied to the designs. At this level, designs look more similar to the actual hardware representation. Data-flow and behavioural abstraction styles are used to implement the large designs.

Building a triplication scheme into digital designs is not a simple task. It takes a lot of time to build and debug the system. It has been also a laborious and error-prone task. Xilinx Inc. with the help of Sandia lab developed Triple Modular Redundancy (XTMR) tool for the design triplication. XTMR works with HDL and synthesis tool to automatically built TMR methodology. It is designed for the SEU and SET immunity for the FPGA devices, which includes Virtex family of reconfigurable FPGAs [11]. It requires two files, with extensions *.ngc (generated by synthesis process) and *.ngo (generated after the NGDBuild process) files, as an input and generates the TMR logic for the design. This tool is vendor specific and only used for Virtex family of FPGA devices. The TMRG (Triple Module Redundancy Generator) tool is developed in Python at CERN research institute, which automates the process of triplication of digital circuits at implementation stage. It is used to mitigate

the SEEs. TMRG works on Verilog code and adds the classical majority voter circuit in TMR approach [12]. Brigham Young University (BYU) developed BL-Tmr tool [13], which works on EDIF-format design. It parses input EDIF file(s) into a net-list data structure and uses the classification information to select circuit structures for triplication. This tool works for partial TMR approach. Another tool, named TLegUp, which automatically generates Triple Modular Redundant designs for FPGAs from C programs using high-level synthesis technology [14]. Recently, some commercial tools such as Synopsys Synplify Premier, and Mentor Precision HiRel are available in the market to implement TMR during the synthesis process [15]. Using commercial tools, the cost of the project increases unnecessarily.

In this work, a tool named RASP-TMR Code Generator (RechnerArchitektur und SystemProgrammierung - Triple Modular Redundancy) is presented, of which the first part is the German name of our department. It takes Verilog HDL design file as an input and generates the synthesizable Verilog code for TMR technique. A new and simple majority voter circuit is also proposed. To validate the operation of the proposed tool, a simulation set-up is created with the help of Xilinx ISE tools and ISim simulator. The TMR operation is validated by injecting bit-flip and stuck at 1/0 faults in the design during the simulation, and it has been observed that the proposed majority voter circuit perfectly masks the errors/failures. This tool, along with its graphical user interface, is developed in MATLAB and it requires the users to provide only Verilog module file and then it automatically generates all the designs necessary to perform TMR. Hence, the RASP-TMR tool is simple, fast and user-friendly. To validate these claims, various benchmark designs are evaluated.

The structure of the paper is as follows: Section II explains the structure of the proposed RASP-TMR in detail, along with the proposed majority voter circuit. Section III presents the results of the RASP-TMR tool for combinational and sequential designs from ISCAS'85, ISCAS'89, and EPFL benchmark designs. Finally, Section IV concludes the paper.

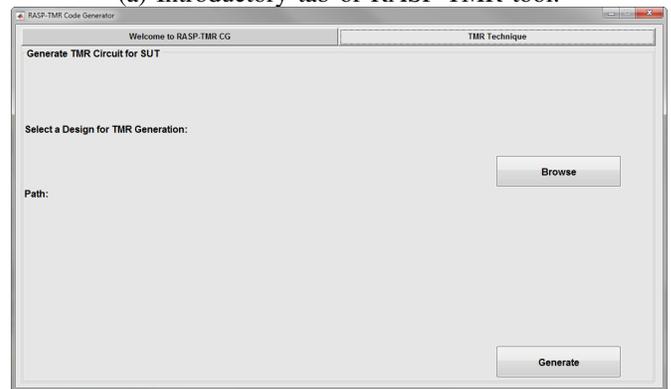
II. PROPOSED RASP-TMR CODE GENERATOR

The RASP-TMR code generator tool, with its graphical user interface, is developed in MATLAB. It is a tabbed based tool and Fig. 1 shows these tabs. The first tab depicts the information about the tool, whereas, the second tab consists of the RASP-TMR tool. The tool consists of 9 functions and 254 lines of code in MATLAB. To provide ease of use, a standalone app is created using MATLAB command `deploytool`. This app is installed by the user on any host machine, having a Windows operating system. This tool not only triplicates the design and generates Verilog top file but also instantiates the designs in the top file. The proposed majority voter circuit is added for each output port of the design.

Fig. 2 depicts the flowchart of the RASP-TMR tool, which includes the series of operations performed by RASP-TMR code generator tool. The RASP-TMR tool accepts a Verilog design file as an input, parses it, obtains tokens (input arguments, output arguments) and makes three copies of it. Module name and the output argument must be changed in order to differentiate from each other and brought them under



(a) Introductory tab of RASP-TMR tool.



(b) RASP-TMR code generator tab.

Fig. 1. GUI of the proposed RASP-TMR code generator.

the top-level file. RASP-TMR also generates the top-level module file which includes input arguments, output arguments, instantiation of all three TMR modules and proposed majority voter circuit for each output port.

A. Top File Structure

The top file consists of different components such as triplication of the module, their instantiations, and a proposed majority voter circuit. Fig. 3 shows the general block diagram of the top file generated by the RASP-TMR. These components are described in the sequel.

1) *Triplication of Module:* The triple modular redundancy requires the triplication of a module. All three modules operate in parallel. If any of a module fails to perform the intended task or results in an error, it is masked by the majority voter circuit. Therefore, when a module is input to the RASP-TMR code generator tool by the user, it parses the design and triplicates it. It should be noted that the module name changes from `c17` to `c17_1`, which denotes the first module of TMR designs. The other two modules are renamed to `c17_2` and `c17_3` respectively. The output ports are changed as shown in Fig. 4. It shows the original Verilog design as an example which is input to the RASP-TMR tool (above) and the output of the tool (below).

2) *Instantiation Generator:* Xilinx ISE design tools provide the built-in instantiation generator for modules available in the design to instantiate one module within another thus creating hierarchy. RASP-TMR has the ability to instantiate the

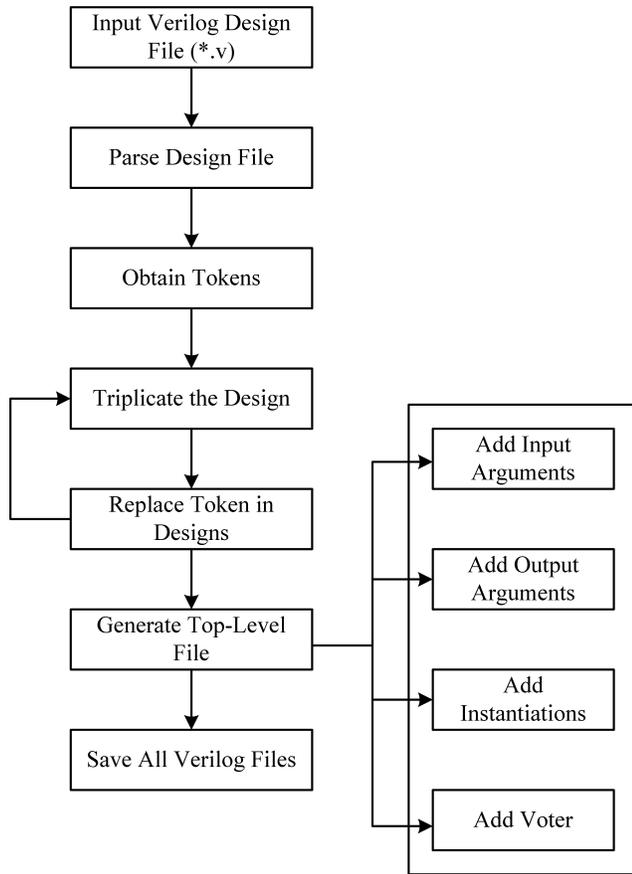


Fig. 2. Flow chart of the RASP-TMR tool.

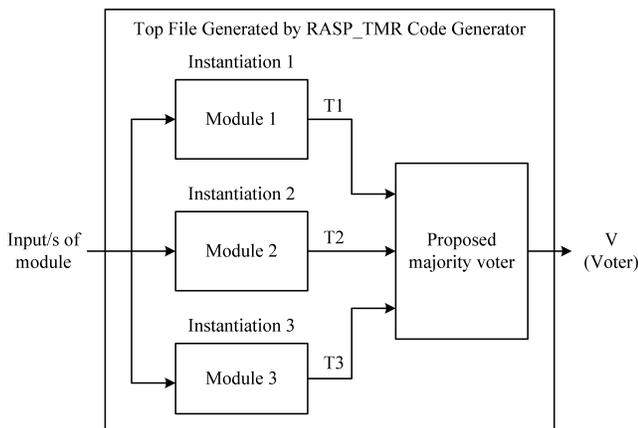


Fig. 3. Structure of top file generated by RASP-TMR tool.

TMR modules in the top file. The function `InstanceGen()` is developed under RASP-TMR tool which generates the instantiations. Instantiation requires the module name, produces instance name and adds the information about the input/output ports. Fig. 5 shows the instantiation for the TMR module 1.

3) *Proposed Majority Voter Circuit*: The TMR scheme involves two-times duplication of the simplex system hardware, with majority voter ensuring correctness provided at least two out of three copies of the system remains operational. Various majority voter designs have been proposed for the last couple

```
// c17
// Original design
module c17 (N1,N2,N3,N6,N7,N22,N23) ;
input N1,N2,N3,N6,N7;
output N22, N23;
wire N10,N11,N16,N19;
nand NAND2_1 (N10, N1, N3);
nand NAND2_2 (N11, N3, N6);
nand NAND2_3 (N16, N2, N11);
nand NAND2_4 (N19, N11, N7);
nand NAND2_5 (N22, N10, N16);
nand NAND2_6 (N23, N16, N19);
endmodule

// c17
// TMR Module 1
module c17_1 (N1,N2,N3,N6,N7,N22_tmr1,
              N23_tmr1) ;
input N1,N2,N3,N6,N7;
output N22_tmr1, N23_tmr1;
wire N10,N11,N16,N19;
nand NAND2_1 (N10, N1, N3);
nand NAND2_2 (N11, N3, N6);
nand NAND2_3 (N16, N2, N11);
nand NAND2_4 (N19, N11, N7);
nand NAND2_5 (N22_tmr1, N10, N16);
nand NAND2_6 (N23_tmr1, N16, N19);
endmodule
```

Fig. 4. Code snippet (original design and modified design).

```
// Instantiate the module
c17_1 inst_tmr1 (
.select1(select1),
.N1(N1),
.N2(N2),
.N3(N3),
.N6(N6),
.N7(N7),
.N22_tmr1(N22_tmr1),
.N23_tmr1(N23_tmr1)
);
```

Fig. 5. Code snippet (instantiation).

of years. Fig. 6 shows the schematic diagram of majority voter circuits. Fig. 6 (a) shows the classical voter design. Some other majority voter logics are proposed by Kshirsagar and Patrikar voter circuit [16], Ban and Naviner majority voter [17], and Balasubramanian and Prasad majority voter circuit [7]. Fig. 6 (b-d) show the schematic of other proposed majority voter circuits, respectively.

In this work, authors proposed another simple way to represent majority voter logic and added to the RASP-TMR tool as shown in Fig. 7. In this figure, T1, T2 and T3 are the inputs of the majority voter circuit. Proposed MVC consists of an AND, OR gates and a multiplexer (2:1). Table I shows the functional verification and correctness of the proposed voter logic.

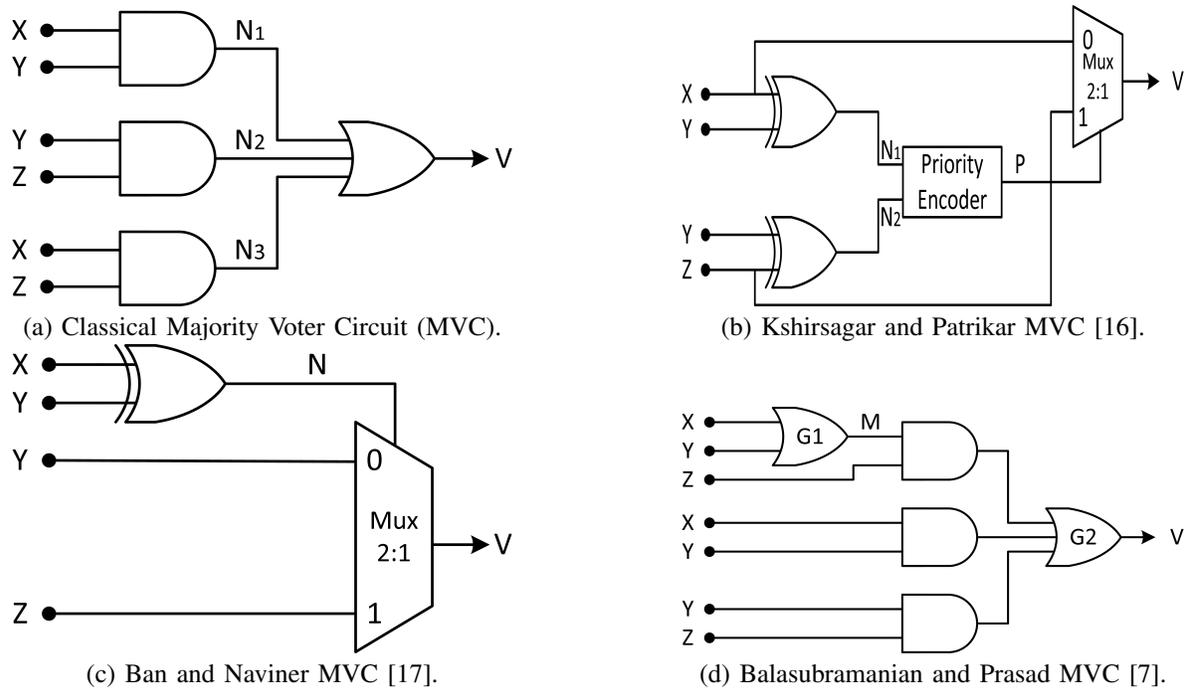


Fig. 6. Various majority voter designs in the literature.

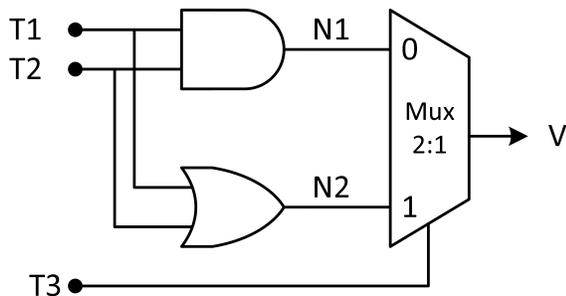


Fig. 7. Proposed majority voter schematic diagram.

TABLE I. TRUTH TABLE VERIFICATION OF PROPOSED MAJORITY VOTER LOGIC FOR TMR

T3	T2	T1	N1	N2	V
0	0	0	0	0	0
0	0	1	0	1	0
0	1	0	0	1	0
0	1	1	1	1	1
1	0	0	0	0	0
1	0	1	0	1	1
1	1	0	0	1	1
1	1	1	1	1	1

III. RESULT AND DISCUSSION

The primary purpose of using a TMR design methodology is to remove all single point of failure from the design. If two redundant modules are simultaneously upset, then the output can not be guaranteed to be correct. In order to develop a TMR logic for a target system at a code level, the code

needs to be modified for each module and also majority voter circuit is included in the design. This is a non-trivial and time-consuming task.

A. Synthesizable Designs

Keeping these points in mind, the RASP-TMR code generator tool is developed for Verilog HDL designs. This tool not only triplicates the target systems but also generates the top file (named *TMR_TopFile*). The components of this file are already described in Section II in detail. At the graphical user interface, the user needs to provide only a synthesizable Verilog HDL design as an input (This verifies our claim about the tool's simplicity and easy to use). A simple benchmark circuit (c17.v) from ISCAS'85 is considered as an example for illustration of the whole process. A project is developed using Xilinx project navigator. The synthesis of the design is performed and generated the RTL schematic of the example shown in Fig. 8. This proves that the designs generated by RASP-TMR tool are synthesizable.

B. Timing Analysis

This tool is fast and automatically generate the TMR designs. To prove this point, authors have generated TMR designs for various benchmark designs, written in Verilog. In this work, ISCAS85, ISCAS89, and EPFL combinational and sequential benchmark designs are considered for their generation of TMR technique and measured the time. Tables II to V show the size of the designs in terms of a number of logic gates. Last columns of these tables show the time taken by this tool in Seconds. It is noted that the TMR approach for the designs, consisting of thousands of gates, are generated in a fraction of the second by the RASP-TMR tool. The design "hypotenuse" is the largest design among them, which consists

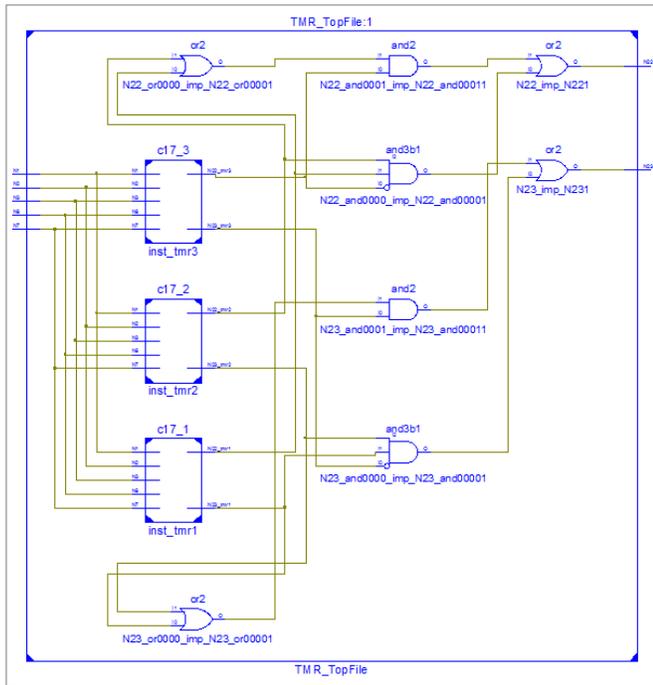


Fig. 8. RTL schematic of c17 circuit with TMR and proposed MVC.

TABLE II. TIME REQUIRED FOR ISCAS'85 COMBINATIONAL BENCHMARK DESIGNS

S. No.	Benchmark circuits	No. of logic gates	Time (Seconds)
1	c17	6	0.096
2	c432	160	0.1048
3	c499	202	0.2124
4	c880	383	0.1109
5	c1355	546	0.7257
6	c1908	880	0.4743
7	c2670	1269	0.6975
8	c3540	1669	0.3416
9	c5315	2307	0.4948
10	c6288	2416	0.3434
11	c7552	3513	0.5694

of 214335 logic gates and RASP-TMR took approximately 390 seconds.

C. Functional Verification of Proposed Majority Voter Circuit

In this tool, authors also proposed a new majority voter logic. In [7], authors describe the procedure to calculate the Fault Mask Ratio (FMR) for various majority voter circuits. According to them, *FMR is the ratio of a total number of correct voter output states divided by the total number of likely internal and/or external faults corresponding to the applied primary inputs.*

For classical majority voter circuit, the value of FMR is 42.86%. Similarly, for other majority voter circuits Fig. 6 (b-d), the values of FMR are 70.83, 50, 75% respectively. In our case,

TABLE III. TIME REQUIRED FOR ISCAS'89 SEQUENTIAL BENCHMARK DESIGNS

S. No.	Benchmark circuits	No. of logic gates/FFs	Time (Seconds)
1	s1238	508/18	0.0643
2	s1423	657/74	0.0700
3	s1488	653/6	0.0779
4	s1494	647/6	0.0784
5	s5378	2779/179	0.3016
6	s9234	5597/211	0.5925
7	s13207	7951/638	1.0365
8	s15850	9772/534	1.2942
9	s35932	16065/1728	3.7814
10	s38417	22179/1636	4.0921
11	s38584	19253/1426	5.0263

TABLE IV. TIME REQUIRED FOR EPFL (ARITHMETIC) BENCHMARK DESIGNS

S. No.	Benchmark circuits	No. of logic gates	Time (Seconds)
1	Adder	1020	0.1985
2	B-shifter	3336	0.4048
3	Divisor	44762	25.265
4	Hypotenuse	214335	390.79
5	Log2	32060	6.8845
6	Max	2865	0.5062
7	Multiplier	27062	5.9782
8	Sine	5412	0.5089
9	Square	24618	5.6014
10	Square-root	18484	3.3004

TABLE V. TIME REQUIRED FOR EPFL (RANDOM/CONTROL) BENCHMARK DESIGNS

S. No.	Benchmark circuits	No. of logic gates	Time (Seconds)
1	Arbieter	11839	1.4639
2	ALU	174	0.0533
3	Coding	693	0.0862
4	Decoder	304	0.1889
5	I2C	1342	0.1995
6	Int-to-float	260	0.0484
7	Memory	46836	32.9611
8	Encoder	978	0.1091
9	Router	257	0.0754
10	Voter	13758	2.1622

the calculated FMR for the proposed majority voter circuit is 50% which is better than the classical approach. Authors simulated the proposed majority voter circuit with all possible combinations of inputs. It can be seen from Fig. 9 that the voter signal has a value logic '0' if the majority of inputs are at logic '0' and vice versa. Hence, this simulation verifies the operation of proposed majority voter logic. It is noted that the

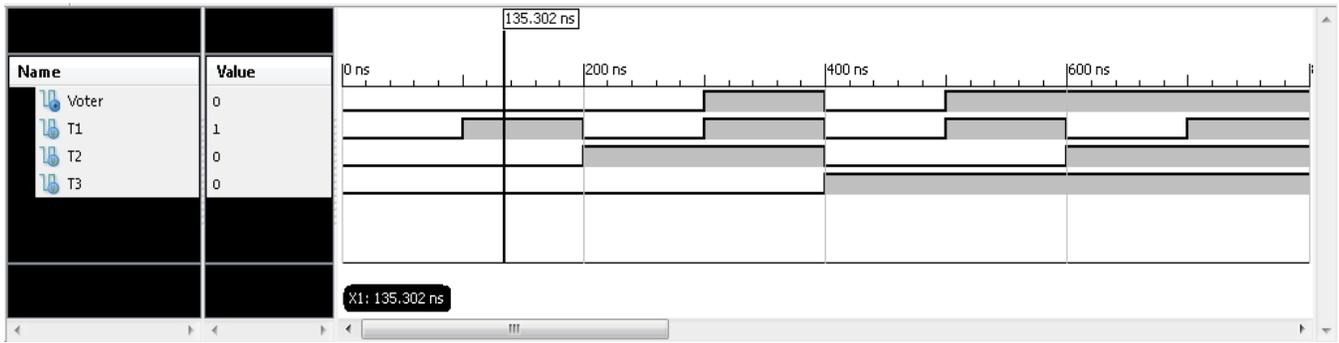


Fig. 9. Simulation results for the validation of proposed majority voter circuit.

shaded area of signals T1, T2, T3, and Voter represent the logic '1' value.

D. Simulation Verification using Fault Injection Technique

Fault injection is a useful technique to test the integrity of the TMR system. In order to test and verify the fault tolerance capability of the target design developed under the RASP-TMR tool, simulation of the design is carried-out with the help of Xilinx ISE Design Suite 13.4 tools and ISim simulator. Xilinx ISE is used to develop the project and writing a test bench, while the simulation is performed by ISim tool. A simulation environmental set-up is created and Fig. 10 shows the block diagram of the set-up. It consists of TMR design and a golden (fault-free) module of the design. Faults are injected in the TMR design in each module. When the faults are activated one by one in each module and different combinations of input are applied and then responses are compared. If the comparator output is logic '0', that means both the golden and TMR outputs are the same and the TMR approach masks the faulty module perfectly. In order to validate the approach, a simple benchmark design from ISCAS'85 combinational circuits has chosen. Authors have injected a total of 12 faults in the c17 benchmark circuit in different locations of the design, 4 faults in each TMR module. The way of injecting faults in the Verilog design code is described in our previous work [3], [18], [19].

In the first instance, the 'fault 0' to 'fault 3' are activated in module 1 of the TMR design using the 2-bit vector `faultIn1`. The input patterns are applied and denoted by input ports (N1, N2, N3, N6, N7). The 'cmp1' signal

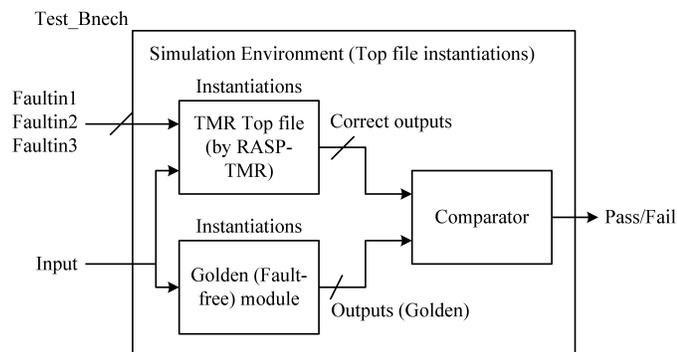


Fig. 10. Simulation environment for the verification of proposed tool.

compares the responses of the golden module with the responses of the TMR approach. Fig. 11 shows that the signal 'cmp1' is continuously at logic '0', which means both the responses are the same. Similarly, the same approach is repeated for the second and third module of the TMR approach and simulation results are shown in Fig. 12 and 13. Hence, it verifies the operations of the TMR approach developed by the RASP-TMR tool and the proposed majority voter circuit.

IV. CONCLUSION

In this work, authors developed a RASP-TMR tool which replicates any combinational and sequential digital designs. The RASP-TMR tool is simple, fast and user-friendly. The user needs to provide a synthesizable Verilog file as an input, and then the tool creates TMR design along with the proposed majority voter logic circuit. TMR design for various combinational and sequential benchmark circuits is generated and evaluated. The results show that the RASP-TMR tool takes less time to generate the synthesizable Verilog code with TMR technique. Benchmark designs are simulated using Xilinx tools to evaluate the features of our proposed tool.

The future work will add more features to the tool such as TMR with Multiple Voters, Duplication With Comparison (DWC) and N-modularity redundancy for the evaluation of the fault-tolerant capability of FPGA-based designs conveniently.

REFERENCES

- [1] G. Corradi, R. Girardey, and J. Becker, "Xilinx tools facilitate development of FPGA applications for IEC61508," in *Adaptive Hardware and Systems (AHS), 2012 NASA/ESA Conference on*, Erlangen, Germany, Jun 2012, pp. 54–61.
- [2] A. R. Khatri, M. Milde, A. Hayek, and J. Börcsök, "Instrumentation Technique for FPGA based Fault Injection Tool," in *5th International Conference on Design and Product Development (ICDPD '14)*, Istanbul, Turkey, Dec 2014, pp. 68–74.
- [3] A. R. Khatri, A. Hayek, and J. Börcsök, *Applied Reconfigurable Computing*, ser. Lecture Notes in Computer Science, V. Bonato, C. Bouganis, and M. Gorgon, Eds. Cham: Springer International Publishing, 2016, vol. 9625. [Online]. Available: <http://link.springer.com/10.1007/978-3-319-30481-6>
- [4] M. Desogus, L. Sterpone, and D. M. Codinachs, "Validation of a tool for estimating the effects of soft-errors on modern SRAM-based FPGAs," in *2014 IEEE 20th International On-Line Testing Symposium (IOLTS)*. Platja d'Aro, Girona, Spain: IEEE, Jul 2014, pp. 111–115. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6873681>

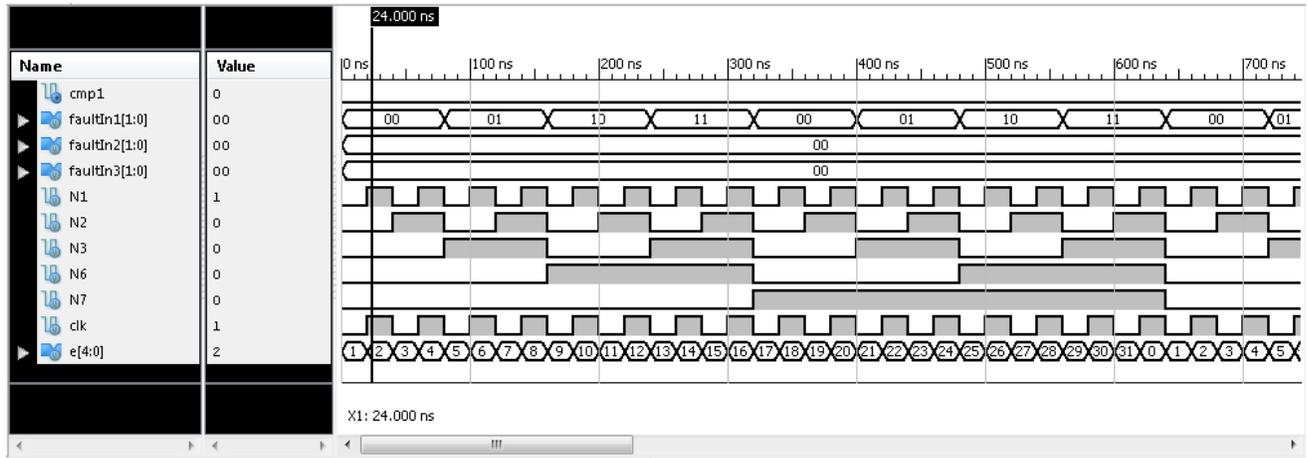


Fig. 11. Simulation results for bit-flip injection in TMR module 1.

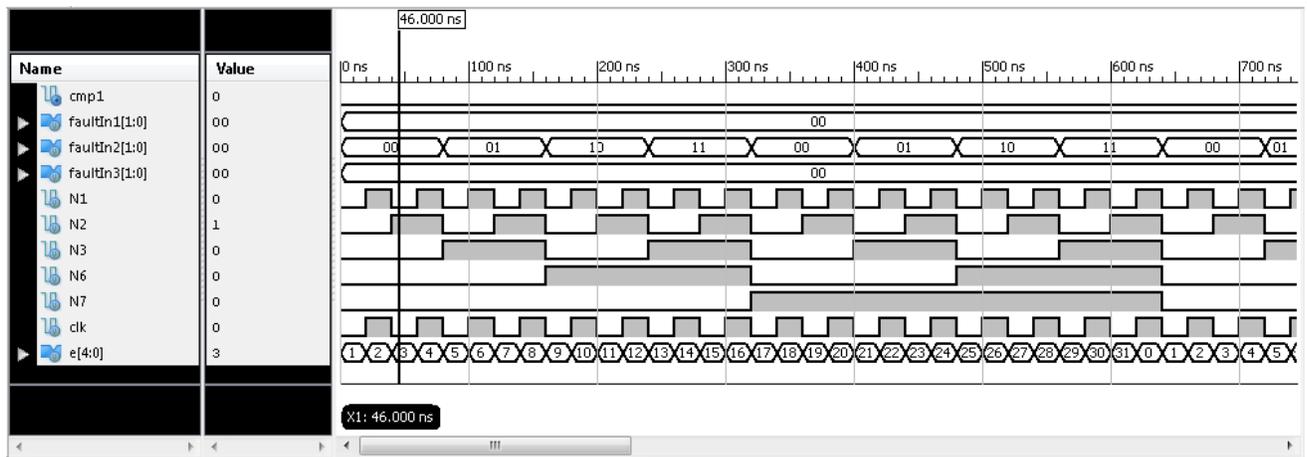


Fig. 12. Simulation results for bit-flip injection in TMR module 2.

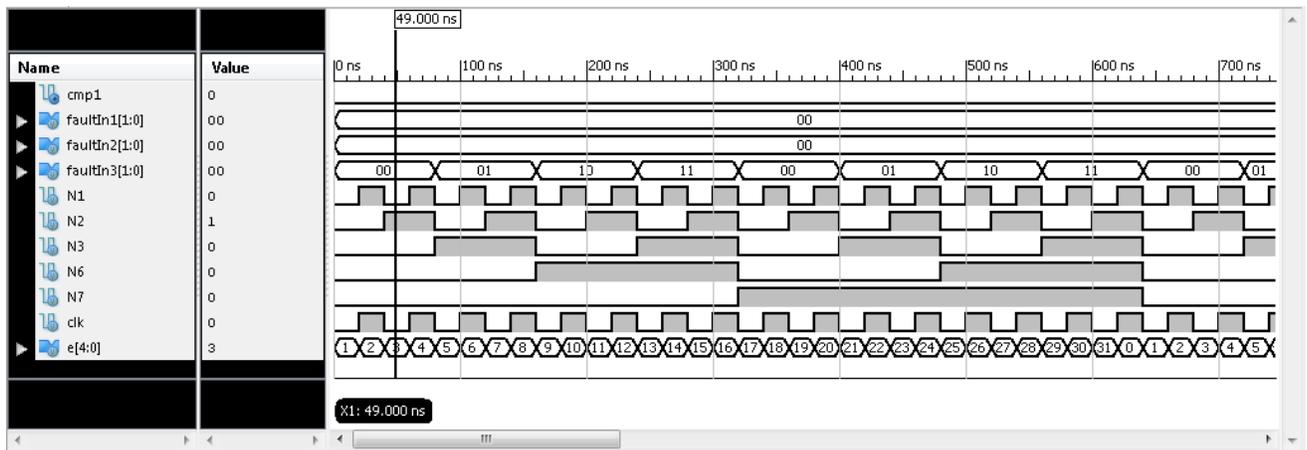


Fig. 13. Simulation results for bit-flip injection in TMR module 3.

[5] C. Frenkel, J.-d. Legat, and D. Bol, "A Partial Reconfiguration-based scheme to mitigate Multiple-Bit Upsets for FPGAs in low-cost space applications," in *2015 10th International Symposium on Reconfigurable Communication-centric Systems-on-Chip (ReCoSoC)*. Bremen: IEEE, Jun 2015, pp. 1–7. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7238095>

[6] W. Xin, "Partitioning Triple Modular Redundancy for Single Event

Upset Mitigation in FPGA," in *2010 International Conference on E-Product E-Service and E-Entertainment*. Henan: IEEE, Nov 2010, pp. 1–4. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5660842>

[7] P. Balasubramanian, K. Prasad, and N. E. Mastorakis, "A Fault Tolerance Improved Majority Voter for TMR System Architectures," *WSEAS Transactions on Circuits and Systems*, vol. 15, pp. 108–122,

2016. [Online]. Available: <http://arxiv.org/abs/1605.03771>
- [8] S. Müller, T. Koal, M. Schölzel, and H. T. Vierhaus, "Timing for Virtual TMR in Logic Circuits," in *IEEE 20th International On-Line Testing Symposium (IOLTS)*, 2014, pp. 190–193.
- [9] S. Di Carlo, G. Gambardella, P. Prinetto, D. Rolfo, P. Trotta, and A. Vallero, "A novel methodology to increase fault tolerance in autonomous FPGA-based systems," in *2014 IEEE 20th International On-Line Testing Symposium (IOLTS)*. Girona, Spain: IEEE, Jul 2014, pp. 87–92. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6873677>
- [10] P. Samudrala, J. Ramos, and S. Katkooi, "Selective triple Modular redundancy (STMR) based single-event upset (SEU) tolerant synthesis for FPGAs," *IEEE Transactions on Nuclear Science*, vol. 51, no. 5, pp. 2957–2969, Oct 2004. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1344451>
- [11] Xilinx, "Xilinx TMRTool User Guide - TMRTool Software Version 13.2," Tech. Rep., 2017. [Online]. Available: https://www.xilinx.com/support/documentation/user_{_}guides/ug156-tmrtool.pdf
- [12] S. Kulis, "Single Event Effects mitigation with TMRG tool," *Journal of Instrumentation*, vol. 12, no. 01, pp. C01082–C01082, Jan 2017. [Online]. Available: <http://stacks.iop.org/1748-0221/12/i=01/a=C01082?key=crossref.0735729cdfa9a0dabd055faa790f2459>
- [13] B. Pratt, M. Caffrey, P. Graham, K. Morgan, and M. Wirthlin, "Improving FPGA Design Robustness with Partial TMR," in *2006 IEEE International Reliability Physics Symposium Proceedings*. IEEE, 2006, pp. 226–232. [Online]. Available: <http://ieeexplore.ieee.org/document/4017162/>
- [14] G. Lee, D. Agiakatsikas, T. Wu, E. Cetin, and O. Diessel, "TLegUp: A TMR Code Generation Tool for SRAM-Based FPGA Applications Using HLS," in *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, no. 3. IEEE, Apr 2017, pp. 129–132. [Online]. Available: <http://ieeexplore.ieee.org/document/7966665/>
- [15] L. A. C. Benites and F. L. Kastensmidt, "Automated design flow for applying Triple Modular Redundancy (TMR) in complex digital circuits," in *2018 IEEE 19th Latin-American Test Symposium (LATS)*. IEEE, Mar 2018, pp. 1–4. [Online]. Available: <https://ieeexplore.ieee.org/document/8349668/>
- [16] R. Kshirsagar and R. Patrikar, "Design of a novel fault-tolerant voter circuit for TMR implementation to improve reliability in digital circuits," *Microelectronics Reliability*, vol. 49, no. 12, pp. 1573–1577, Dec 2009. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0026271409003114>
- [17] T. Ban and L. A. Naviner, "A simple fault-tolerant digital voter circuit in TMR nanoarchitectures," in *Proceedings of the 8th IEEE International NEWCAS Conference 2010*. IEEE, Jun 2010, pp. 269–272. [Online]. Available: <http://ieeexplore.ieee.org/document/5603933/>
- [18] A. R. Khatri, A. Hayek, and J. Börcsök, "ATPG method with a hybrid compaction technique for combinational digital systems," in *2016 SAI Computing Conference (SAI)*. London, UK: IEEE, Jul 2016, pp. 924–930. [Online]. Available: <http://ieeexplore.ieee.org/document/7556091/>
- [19] A. R. Khatri, A. Hayek, and J. Börcsök, "Validation of the Proposed Fault Injection, Test and Hardness Analysis for Combinational Data-flow Verilog HDL Designs under the RASP-FIT Tool," in *IEEE 16th Int. Conf. on Dependable, Autonomic & Secure Comp., 16th Int. Conf. on Pervasive Intelligence & Comp., 4th Int. Conf. on Big Data Intelligence & Comp., and 3rd Cyber Sci. & Tech. Cong.* Athens, Greece: IEEE Comput. Soc, Aug 2018, pp. 544–551.

Design of Linear Time Varying Flatness-Based Control for Single-Input Single-Output Systems

Marouen Sleimi

Research Laboratory in Automatic
Control - LA.R.A

University of Tunis EL Manar
National Engineering School of Tunis,
BP-36, le Bilvedere 1002 Tunis

Mohamed Ben Abdallah

Higher Institute of Technology
Study of Rades,

BP 172, 2098, Radès Médina, Tunisia

Mounir Ayadi

Research Laboratory in Automatic
Control - LA.R.A

University of Tunis EL Manar
National Engineering School of Tunis,
BP-36, le Bilvedere 1002 Tunis

Abstract—In this paper, the control of linear discrete-time Varying Single-Input Single-Output systems is tackled. By using flatness theory combined with a dead-beat observer, a two degree of freedom controller is designed with high performances in terms of trajectory tracking. The aim of this work is to avoid the choice of closed loop poles in linear discrete-time varying framework which build a very serious problem in system control. The effectiveness of this control law is highlighted by simulation results.

Keywords—Flatness theory; discrete-time systems; linear time varying; single-input single-output; dead-beat observer; two degree of freedom controller

I. INTRODUCTION

The theory of linear time-invariant systems gives a wide range of design methods and solutions to control problems including all kinds of techniques such as state feedback controllers and observers, Kalman filters, H_2 control and H_∞ control. Since then, researchers have kept on making efforts to extend the previous systems control approaches to time-varying systems and sampled-data systems.

LTV (Linear Time Varying) systems are of a great interest because of the fact that time invariant nonlinear systems can be approximated by LTV systems around desired trajectories after their linearization. However, a fundamental part in the study of LTV systems is insured by the state transition matrix noted ϕ , which can be computed as the sum of the Peano-Baker series. We can note here that not all arguments and assumptions used of time-invariant systems are useful in time-varying framework. To extend them, time-varying systems are examined carefully on their controllability and stability.

Moreover, many control design approaches use LTV systems, one of the most important way of systems control is SMC (Sliding mode control) which is very used in the case of dynamic uncertain systems [1]. Furthermore, we can evaluate Back Stepping method which used to aim the objective of stabilizing nonlinear dynamic systems leading to an LTV system after linearization [2]. Besides, we find fuzzy control operating nonlinear systems to make nonlinear controllers via the use of heuristic information [3]. Among these control design approaches, flatness-based control remains the most suitable method in trajectory tracking, then in the rest of the paper, we are interested to this kind of controller in the LTV case.

Previously, it is shown that flatness property considerably simplifies the 2DOF (Two Degree of Freedom) controllers design for continuous-time SISO (Single-Input Single-Output) systems for LTV framework [4], [5]. In these works, the main feature of this flatness approach for 2DOF controllers design, using flatness-based control and dead-beat observer, is to avoid the choice of closed loop poles and no need to solve diophantine equation any more. In this design the closed loop dynamics are related to the chosen tracking dynamics.

This approach was extended to discrete-time framework for LTI SISO [6], [7] and LTI MIMO (Multiple-Input Multiple-Output) systems [8], [9], [10]. This paper extend the previous approaches to deal with LTV flatness-based control in SISO discrete-time framework.

This paper is organized as follows: In Section II, some preliminaries are presented. Then in Section III, the definition of the canonical controllable form in discrete-time SISO LTV framework is given. Moreover, in Section IV, the new approach of control design is developed. In Section V, effectiveness of this control law is illustrated using an academic discrete-time SISO system.

In the following, we will develop the paper in a discrete-time formulation, using the shift forward operator q and the delay operator q^{-1} .

II. PRELIMINARIES

Introducing a given problem concerning LTV systems, start generally by results given by LTI techniques and trying to adapt it in the new context. Flatness-based control is introduced and developed by Fliess and co-researchers [11] and used by many authors, firstly in the LTI framework then in the LTV one.

This work deals with flatness-based control in LTV discrete-time framework. So to aim this objective a canonical controllable form must be built and is exploited in the proposed control law design.

A. Canonical controllable form for LTV discrete-time SISO systems

Canonical forms are widely used in control theory. In this section, the discrete-time LTV controllable canonical form is presented [12].

Considering the following discrete-time LTV system given by:

$$\begin{cases} x_{k+1} = A_k x_k + B_k u_k \\ y_k = C_k x_k \end{cases} \quad (1)$$

where x_k is a n dimensional vector, A_k is a $(n \times n)$ matrix, B_k and C_k are a m dimensional vectors. If the given system is uniformly N-step controllable, then there exist a transformation applied to the state given by:

$$\bar{Z}_k = \bar{T}_k x_k \quad (2)$$

where \bar{T}_k is a Lyapunov transformation matrix, presented in Appendix A, obeying the conditions below.

- \bar{T}_k is defined each sample time.
- \bar{T}_k and \bar{T}_{k+1} are bounded each sample time.
- There exist a constant m where:

$$0 < m < \det(\bar{T}_k), \quad k > 0$$

This algorithm of controllable form leads to a new state vector equation written as following:

$$\begin{cases} Z_{k+1} = \bar{A}_k Z_k + \bar{B}_k \bar{u}_k \\ y_k = \bar{C}_k Z_k \end{cases} \quad (3)$$

where \bar{A}_k , \bar{B}_k and \bar{C}_k are given by the new canonical controllable form for LTV systems, such as:

$$\bar{A}_k = \bar{T}_{k+1} A_k \bar{T}_k^{-1} \quad (4)$$

$$= \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & 0 \\ 0 & \dots & \dots & 0 & 1 \\ -\gamma_0(k) & -\gamma_1(k+1) & \dots & \dots & -\gamma_{n-1}(k+n-1) \end{bmatrix} \quad (5)$$

$$\bar{B}_k = \bar{T}_{k+1} B_k = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad (6)$$

$$\bar{C}_k = C_k \bar{T}_{k+1} = (\alpha_{1k} \quad \dots \quad \alpha_{(n-1)k}) \quad (7)$$

B. Flatness Properties

Flatness can parametrize, in a very simple way, the dynamic aspect of a given system based on highlighting some endogenous fundamental variables called flat outputs. In fact, the state, the input and the output using the flat output can be written as follows:

$$x_k = f_1(z_k, z_{k+1}, \dots, z_{k+n-1})^T \quad (8)$$

and

$$u_k = f_2(z_k, z_{k+1}, \dots, z_{k+m-1})^T \quad (9)$$

$$y_k = g_k(x_k, u_k)^T \quad (10)$$

As known, on the first hand, the state, the input vector and the output vector in continuous time framework are defined as a successive derivatives of the flat output, on the other hand, in discrete-time framework it is considered as a phase advance sequence of the flat output.

Marlait et al. [13] results confirm that discrete-time LTV system is the equivalent of a controllable LTI system and noting that flatness of a given one is strongly related to the uniform controllability. To build up flatness-based control it is necessary that the considered system is flat, so firstly its controllability [12], [14] must be checked out. If the system is uniformly N-step controllable then we can apply the control law. Let's consider the plant given by (1) and the transformation given by (2).

Noting that z_k the first element of the state vector Z_k , then from (3) it's clear that Z_k can be written as [11]:

$$Z_k = [z_k \quad z_{k+1} \quad \dots \quad z_{k+n-1}]^T \quad (11)$$

where z_k is the Bronovsky output.

The expression of the input u_k is a function of z_k , and its forward terms:

$$u_k = z_{k+n} + \gamma_{n-1}(k+n-1)z_{k+n-1} + \dots + \gamma_1(k+1)z_{k+1} + \gamma_0(k)z_k \quad (12)$$

The output y_k which depends on z_k is written as follows:

$$y_k = \bar{C}_k Z_k = \sum_{i=0}^{n-1} \alpha_{ik} z_{k+i} \quad (13)$$

z_k can be considered as a flat output of the discrete LTV system. From the canonical form given previously flatness-based control for SISO LTV systems will be developed in the next section.

III. FLATNESS-BASED CONTROL FOR SISO LTV SYSTEMS

A. Flatness-based Control

Following (11) the flatness-based control law is given by:

$$u_k = z_{k+n}^d + \sum_{i=0}^{n-1} \kappa_i (z_{k+i}^d - z_{k+i}) + \sum_{i=0}^{n-1} \gamma_i (k+i) z_{k+i} \quad (14)$$

κ_i are derived from the following polynomial:

$$K(q) = q^n + \sum_{i=0}^{n-1} \kappa_i q^i \quad (15)$$

which must be shur [15], [16]. To build this control it is necessary to estimate the state vector Z_k using an observer [17]. In this paper the dead-beat observer is used in the design strategy leading to a 2DOF controller. The structure of the flatness-based control is given by the following expression:

$$u_k = K(q)z_k^d + \Lambda_k Z_k \quad (16)$$

with:

$$\Lambda_k = \begin{bmatrix} \gamma_{0k} - \kappa_0 \\ \vdots \\ \gamma_{i(k+i)} - \kappa_i \end{bmatrix} \quad (17)$$

The control law schema is shown in Fig. 1.

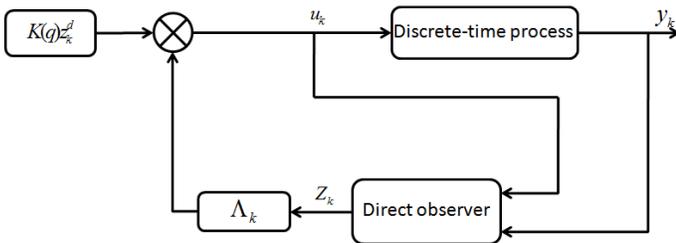


Fig. 1. Control law design.

B. Dead-beat Discrete-time Varying Observer Design

Successive iterations of the output equation in (3) to the order $(n - 1)$ [18], give:

$$\begin{pmatrix} y_k \\ y_{k+1} \\ \vdots \\ y_{k+n-1} \end{pmatrix} = L_k Z_k + M_k \begin{pmatrix} u_k \\ u_{k+1} \\ \vdots \\ u_{k+n-2} \end{pmatrix} \quad (18)$$

with:

$$L_k = \begin{pmatrix} \bar{C}_k \\ \bar{C}_{(k+1)} \bar{A}_k \\ \bar{C}_{(k+2)} \bar{A}_{k+1} \bar{A}_k \\ \vdots \\ \bar{C}_{(k+n-1)} \bar{A}_{(k+n-2)} \dots \bar{A}_k \end{pmatrix} \quad (19)$$

and:

$$M_k = \begin{bmatrix} M_{1,k} & 0 & \dots & 0 \\ M_{2,k} & M_{1,k+1} & \dots & \vdots \\ M_{3,k} & M_{2,k+1} & \dots & \vdots \\ \vdots & \vdots & \dots & 0 \\ M_{n-1,k} & M_{n-2,k+1} & \dots & M_{1,k+n-1} \end{bmatrix} \quad (20)$$

where:

$$M_{1,k} = \bar{C}_{k+1} \bar{B}_k \quad (21)$$

and:

$$M_{i,k} = \bar{C}_{k+i-1} \bar{A}_{k+i} \bar{B}_k \quad (22)$$

Noting that:

$$Y_k = (y_k \quad \dots \quad y_{k+n-1})^T \quad (23)$$

and:

$$U_k = (u_k \quad \dots \quad u_{k+n-2})^T \quad (24)$$

Equation (18) can be then written as:

$$Y_k = L_k Z_k + M_k U_k \quad (25)$$

Supposing that the system is uniformly N-step observable then the rank of L_k is n for all k . Noting that L_k is the observability matrix, so:

$$Z_k = L_k^{-1} Y_k - L_k^{-1} M_k U_k \quad (26)$$

In a second hand:

$$Z_k = q^{-1} (\bar{A}_k Z_k + \bar{B}_k u_k) \quad (27)$$

After a second iteration:

$$Z_k = q^{-1} (A_k q^{-1} (A_k Z_k + B_k u_k) + B_k u_k) \quad (28)$$

The iteration of this expression to the order l gives:

$$Z_k = \phi_l q^{-l} Z_k + \sum_{i=1}^{l-1} \phi_i \bar{B}_{k-(i+1)} u_{k-(i+1)} + \bar{B}_{k-1} u_{k-1} \quad (29)$$

with:

$$\phi_l = \prod_{j=1}^l \bar{A}_{k-j} \quad (30)$$

$$\phi_0 = I_n \quad (31)$$

I is the identity matrix and ϕ_l is $(n \times n)$ -dimensional matrices. By replacing Z_k in the right side of (29) by the expression in (26), Z_k can be written as follows:

$$Z_k = \phi_l q^{-l} (L_k^{-1} Y_k - L_k^{-1} M_k U_k) \quad (32)$$

$$+ \sum_{i=1}^{l-1} \phi_i \bar{B}_{k-(i+1)} u_{k-(i+1)} + \bar{B}_{k-1} u_{k-1}$$

then:

$$Z_k = \phi_l L_{k-l}^{-1} \begin{pmatrix} y_{k-l} \\ \vdots \\ y_k \end{pmatrix} - \phi_l L_{k-l}^{-1} M_{k-l} \begin{pmatrix} u_{k-l} \\ \vdots \\ u_{k+n-2-l} \end{pmatrix} \quad (33)$$

$$+ \sum_{i=1}^{l-1} \phi_i \bar{B}_{k-(i+1)} u_{k-(i+1)} + \bar{B}_{k-1} u_{k-1}$$

As Z_k is written as:

$$Z_k = \phi_l L_{k-l}^{-1} \begin{pmatrix} y_{k-l} \\ \vdots \\ y_k \end{pmatrix} - \phi_l L_{k-l}^{-1} M_{k-l} \begin{pmatrix} u_{k-l} \\ \vdots \\ u_{k+n-2-l} \end{pmatrix} \quad (34)$$

$$+ (\phi_{l-1}B_{k-l} \quad \dots \quad \phi_1B_{k-2} \quad \phi_0B_{k-1}) \begin{pmatrix} u_{k-l} \\ \vdots \\ u_{k+n-2-l} \end{pmatrix}$$

we obtain:

$$Z_k = \phi_l L_{k-l}^{-1} Y_{k-l} - [\phi_l L_{k-l}^{-1} M_{k-l} - \Psi] U_{k-l} \quad (35)$$

Supposing that $l = n - 1$, the expression of Z_k is then given as follows:

$$Z_k = \phi_{(n-1)} L_{k-(n-1)}^{-1} Y_{k-(n-1)} \quad (36)$$

$$- [\phi_l L_{k-(n-1)}^{-1} M_{k-(n-1)} - \Psi] U_{k-(n-1)}$$

with:

$$\Psi = [\phi_{(n-1)-1} B_{k-(n-1)} \quad \dots \quad \phi_1 B_{k-2} \quad \dots \quad \phi_0 B_{k-1}] \quad (37)$$

C. 2-DOF controller using flatness based control

By replacing Z_k by its expression of (36) in the expression of the flatness-based control law given by (16).

The flatness-based control can be then written in the following form:

$$S(k, q^{-1}) u_k = K(q) z_k^d - R(k, q) y_k \quad (38)$$

with:

$$S(k, q^{-1}) = (1 + \Lambda_k [\phi_{n-1} L_{k-n-1}^{-1} M_{k-n-1} - \Psi]) \Pi^* \quad (39)$$

and:

$$R(k, q) = \Lambda_k \phi_l L_{k-n-1}^{-1} \Pi \quad (40)$$

where:

$$\Pi = (q^{-n+1} \quad \dots \quad 1)^T \quad (41)$$

and:

$$\Pi^* = (q^{-n+1} \quad \dots \quad q^{-1})^T \quad (42)$$

The final form of the controller allows us to obtain a 2DOF controller in LTV framework without need to define any observation dynamics and without resolution of the diophantine equation.

IV. ILLUSTRATIVE NUMERICAL EXAMPLE

A. Considered model

Let's consider the following discrete-time system defining by the matrices:

$$A_k = \begin{bmatrix} 0 & e^{-kT} \\ 1 & e^{-kT} \end{bmatrix} \quad (43)$$

$$B_k = \begin{bmatrix} 1 \\ e^{-(k+1)T} \end{bmatrix}, \quad C_k = [0 \quad 1]$$

This system is an academic second order model, with single input and single output used to highlight the effectiveness of the discrete-time flatness-based control approach in LTV SISO case. Noting that, k is the iteration rank and $T = 0.5s$ is the simple time.

The open loop step response is given by Fig. 2.

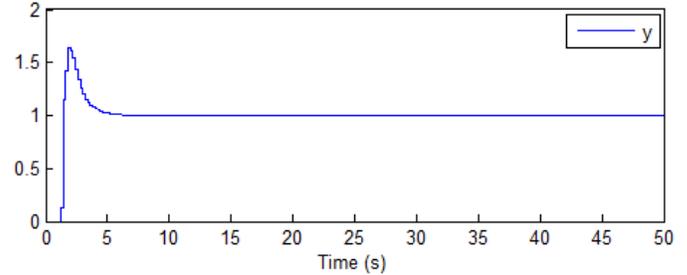


Fig. 2. Control law design.

B. Desired Trajectory

To achieve the implementation of this control law, firstly the definition of the trajectory is needed to be tracked by the considered system.

The aim is to force any system to track the given trajectory, so it's necessary to choose the reference trajectory, then we can determine correctly the endogenous parameters satisfying this objective with a correct dynamics. Let's consider the desired flat output defined as follows:

$$z_d(kT) = z_{in} + \left(21 \left(\frac{kT}{50} \right)^5 - 35 \left(\frac{kT}{50} \right)^6 + 15 \left(\frac{kT}{50} \right)^7 \right) \times (z_{fin} - z_{in}) \quad (44)$$

z_{in} and z_{fin} are the initial and final values of the flat output that is determined endogenously from the initial and final values of the system outputs. The trajectory presented in the discrete-time is used as a reference for the flatness-based control using dead-beat observer.

The discrete-time desired trajectory is represented in Fig. 3.

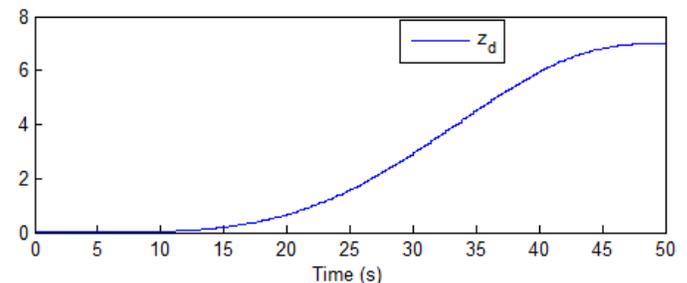


Fig. 3. Desired trajectory for the discrete-time flat output.

C. Control design

From (14), the control law is given by:

$$u_k = z_{k+2}^d + k_1 z_{k+1}^d + k_0 z_k^d + (\gamma_0(k) - k_0) z_k + (\gamma_1(k) - k_1) z_{k+1} \quad (45)$$

After applying the controllable canonical form, the system matrices are given as follows:

$$\bar{A}_k = \begin{bmatrix} 0 & 1 \\ -\gamma_0(k) & -\gamma_1(k) \end{bmatrix} \quad (46)$$

$$\bar{B}_k = \begin{bmatrix} 1 \\ e^{-(k+1)T} \end{bmatrix}, \quad \bar{C}_k = [0 \quad 1]$$

with:

$$\gamma_0(k) = \frac{e^{(k+3)} - e^{(3)} + e^{(3k+1)}}{e^{(2k+1)} + e^{(4k+1)} - e^k} \quad (47)$$

$$\gamma_1(k) = \frac{f_1 + f_2}{f_3} \quad (48)$$

and:

$$\Lambda_k = \begin{bmatrix} \frac{f_4}{f_5} \\ \frac{f_6}{f_7} \end{bmatrix} \quad (49)$$

where:

$$f_1 = e^3 \times e^k - e^1 - e^{2k} \times e^1$$

$$f_2 = e^{2k} \times e^4 + e^{3k} \times e^3$$

$$f_3 = e^{2k} \times e^2 - e^k + e^{4k} \times e^4$$

$$f_4 = 100(e^{k+3} - e^3 + e^{3k+1}) + 7(e^{2k+1} + e^{4k+1} - e^k)$$

$$f_5 = 100(e^{2k+1} + e^{4k+1} - e^k)$$

$$f_6 = \frac{e^{k+4} + e^3 + e^{2k+3}}{e^{k+2} + e^{3k+4} - 1} - e^{k+1}(e^{-2k} - 1)$$

$$f_7 = e^{k+2} + e^{3k+4} - 2.65$$

The transition matrix is:

$$\phi_1 = \bar{A}_{k-1}, \quad \phi_0 = 1$$

and:

$$\Psi = [\bar{A}_{k-1} \bar{B}_{k-2} \quad \bar{B}_{k-1}]$$

As noted previously, this control law can lead to a 2DOF controller with the following parameters:

$$S(k, q^{-1}) = (1 + \Lambda [\phi_1 L_{k-1}^{-1} M_{k-1} - \Psi]) \Pi^* \quad (50)$$

$$R(k, q) = \Lambda \phi_1 L_{k-1}^{-1} \Pi \quad (51)$$

with:

$$S(k, q^{-1}) = \left[1 - \left(\frac{s_1 \times s_2}{s_3} \right) - \frac{s_4}{s_5} \right] q^{-1} \quad (52)$$

$$R(k, q) = [r_1, r_2] \times \begin{bmatrix} q^{-1} \\ 1 \end{bmatrix} \quad (53)$$

where:

$$s_1 = \frac{e^{k+1} \cdot (e^{2k+1} + 1) - (e^{k+4} + e^{k+1} + 20)}{e^{k+2} + e^{3k+4} - 1} + 0.6$$

$$s_2 = e^{2k} - e^{k+3} + e^2 + e^{2k+1} - e^{3k+1} - e^{2k+3} + e^{4k+1} - e^k$$

$$s_3 = e^{2k+1} + e^{4k+1} - e^k$$

$$s_4 = 100(e^{k+3} + e^{3k+1}) + 13(e^{2k+1} + e^{4k+1} - e^k) - 2$$

$$s_5 = 100(e^{2k+1} + e^{4k+1} - e^k)$$

and:

$$r_1 = \frac{e^{k+3} + 0.13e^{2k+1} + e^{3k+1} + 0.13e^{4k+1} + 0.13e^k - 20}{e^{2k+1} + e^{4k+1} - e^k}$$

$$r_2 = \frac{(r_3 \times r_4 \times r_5)}{r_6}$$

$$r_3 = e^{3k+1}(e^{3k+2} + e^k - 1)$$

$$r_4 = e^{k+5} - e^6 + e^{3k+1}$$

$$r_5 = \frac{e^{k+1}(e^{-2k} + 1) - e^{k+4} + e^{2k+3} + 20}{e^{k+2} + e^{3k+4} - 1} + 0.6$$

$$r_6 = (e^{5k} - e^{2k+5} + e^{3k+4})(e^{4k} - e^{k+2} + e^{2k+2})$$

A 2-DOF controller in LTV framework is obtained without defining any observation dynamics, the effectiveness of this control method is shown in the next section.

D. Simulation Results

After applying the control law signal shown in Fig. 4 to the considered system, the effectiveness of this method is proven and the system output follows the desired flat output with an error which tends to zero as shown in Fig. 6. Both the system output and the desired flat output are represented in Fig. 5.

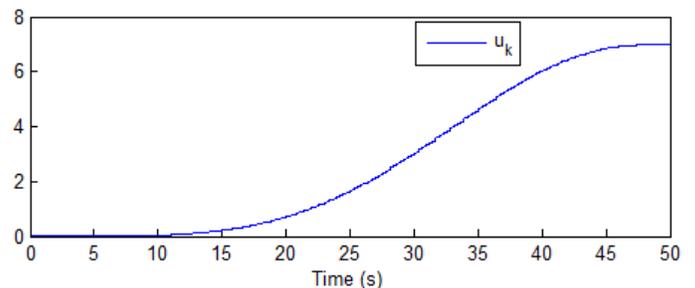


Fig. 4. The control signal u_k .

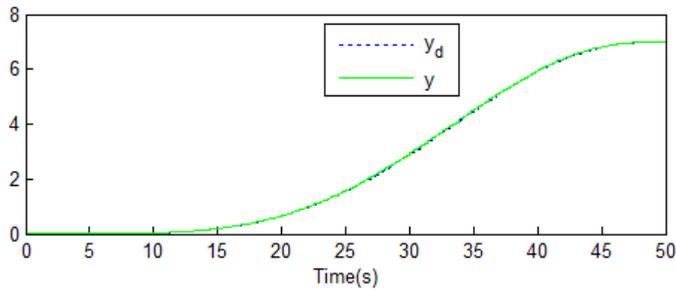


Fig. 5. System output and desired flat output: y and y_d .

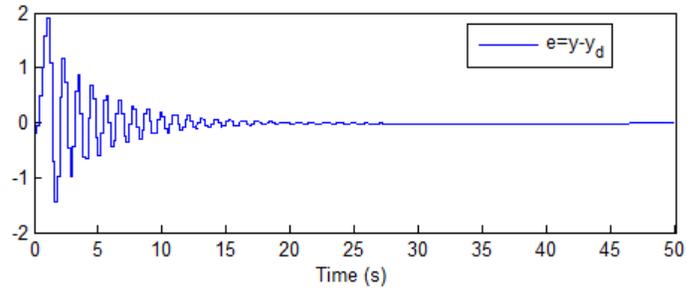


Fig. 9. Tracking error for the second initial condition.

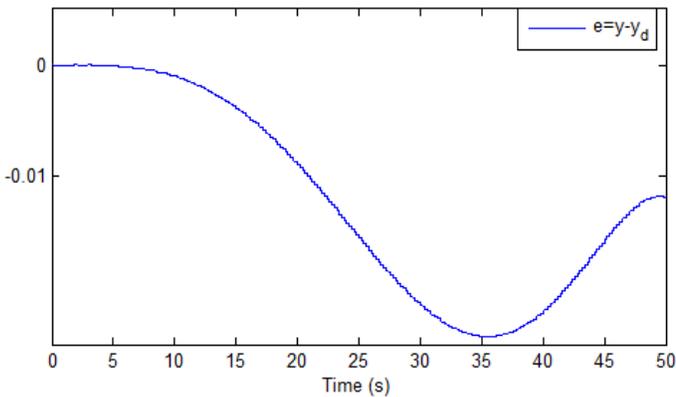


Fig. 6. Tracking error for the system output.

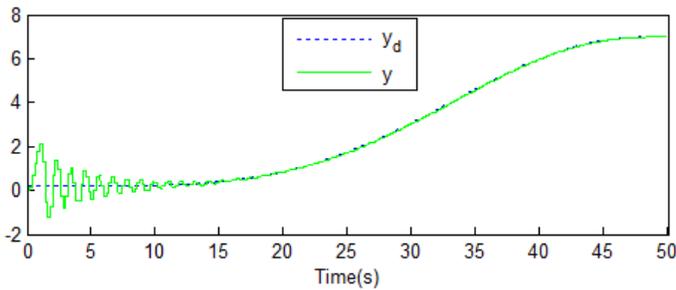


Fig. 7. System output and desired flat output: y and y_d with the first initial condition $\delta y_0 = 0.2$.

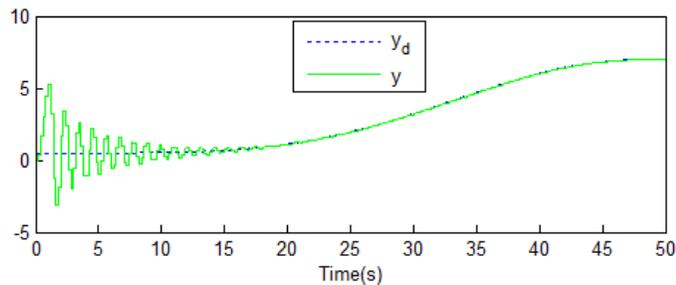


Fig. 8. System output and desired flat output: y and y_d with the second initial condition $\delta y_0 = 0.5$.

When applying two different initial conditions value to the desired trajectory, we obtain the results shown in Fig. 7 and 8

with error for the first one given by Fig. 9.

The linearization around a given trajectory applied to a nonlinear system leads to a linear time-varying system. The real time simulation of such method poses a very important field of study because of the strong symbolic calculations leading to a large size of polynomial matrices. Moreover, The robustness of such method touching the time-varying perturbations, the varying parameters and the practical implementation are prospects which will be carried out in the future works.

V. CONCLUSION

In this paper, a discrete-time control law using the flatness property with trajectory tracking is proposed. Many works which have used the flatness theory to achieve a trajectory tracking by solving bezout equation and carrying out a pole placement. Several difficulties gone through during the design of those methods. Contrary, with a quite few calculations and avoiding the above-mentioned problems our developed approach seems more effective. The design of this control law coupled with a dead-beat observer leads to a 2DOF controller without need to solve diophantine equation. Such a controller based on flatness implies the use of a dead-beat observer, leading to a direct interpretation of the closed loop poles, which are naturally constituted by the poles of the the tracking polynomial $K(q)$. This control law method shows good performances in terms of trajectory tracking in the case of LTV discrete-time systems.

APPENDIX

The matrix \bar{T}_k is called the transformation matrix, its the tool that transforms any system to its controllable form [12]. In the LTV discrete-time case this matrix changes over time at each sample until convergence. The plant (1) becomes:

$$\begin{cases} Z_{k+1} = \bar{T}_{k+1} A_k \bar{T}_k^{-1} Z_k + \bar{T}_{k+1} B_k U_k \\ y_k = C_k \bar{T}_k^{-1} Z_k \end{cases} \quad (54)$$

with $\bar{A}_k = \bar{T}_{k+1} A_k \bar{T}_k^{-1}$, $\bar{B}_k = \bar{T}_{k+1} B_k$ and $\bar{C}_k = C_k \bar{T}_k^{-1}$. In the following, we give the steps to obtain the controllable canonical form. Consider the following vector sequence for $i = 1 \dots n$.

$$\begin{cases} \mathcal{R}_0 = B_k, & i = 0 \\ \mathcal{R}_i = A_k \mathcal{R}_{i-1}(k+1), & i = 1 \dots n \end{cases} \quad (55)$$

the controllability matrix is then written as:

$$\mathcal{R}(k) = (\mathcal{R}_0(k) \quad \dots \quad \mathcal{R}_{n-1}(k)) \quad (56)$$

if $\text{rank}(R(k)) = n$ and to be uniformly N-step controllable means that, for all k , the system is controllable on the interval $[k - N + 1, K + 1]$. Then the effect of this variable change results in the following equation [12]:

$$\bar{\mathcal{R}}_i(k) = \bar{T}_{k+1} \mathcal{R}_i(k) \quad (57)$$

then:

$$\mathcal{R}_c(k) = \bar{T}_{k+1} \mathcal{R}(k) \quad (58)$$

When the pair (\bar{A}_k, \bar{B}_k) is deduced by the variables change of the pair (A_k, B_k) , and if $\mathcal{R}(k)$ is non-singular, then the following form is obtained [7]:

$$\bar{T}_{k+1} = \mathcal{R}_c(k) \mathcal{R}^T(k) (\mathcal{R}(k) \mathcal{R}^T(k))^{-1} \quad (59)$$

otherwise we have the following form:

$$\bar{T}_{k+1} = \mathcal{R}(k) \mathcal{R}_c^{-1}(k) \quad (60)$$

The condition of singularity of the matrix \mathcal{R}_k is a controllability criterion in the discrete-time case. This property is preserved by the variables change [12]. The algorithm for constructing the controllable form is then:

- Construction of $\gamma(k)$ vector

$$\gamma(k) = -\mathcal{R}^{-1}(k) \mathcal{R}_n(k) = (\gamma_0(k) \quad \dots \quad \gamma_{n-1}(k))^T \quad (61)$$

- Transformation \bar{T}_{k+1} then \bar{T}_k which puts the pair (\bar{A}_k, \bar{B}_k) as in (5), (6) and (7).

REFERENCES

- [1] Utkin, V.I., Variable structure systems with sliding mode, *IEEE Transactions on Automatic control*, vol. 22, n. 2, pp.212–222, 1977.
- [2] Carrol, J.J., Schneider, M., and Dawson, D.M., Integrator backstepping techniques for the tracking control of permanent magnet brush Dc motors, *In conference Record of the 1993 IEEE Industry Applications Society Annual Meeting*, pp.663–671, 1993.
- [3] Passino, K.M. and Yurkovich, S., *Fuzzy control*. Addison-wesley Longman. INC, 1998.
- [4] ROTELLA, F., CARRILLO, F. J. and AYADI, M., Polynomial controller design based on flatness, *Kybernetika*, vol. 38, no. 5, pp. 571–584, 2002.
- [5] BEN ABDALLAH, M., AYADI, M., ROTELLA, F. et BENREJEB, M., Régulateurs polynomiaux par platitude pour la commande des systèmes non stationnaires, *Conférence Internationale Francophone de l'Automatique*, CIFA 2012, Grenoble.
- [6] Ayadi, M., Contributions à la commande des systèmes linéaires plats de dimension finie, *Thèse de Doctorat*, Institut National Polytechnique de Toulouse, Tarbes 2002.
- [7] Rotella, F., Carrillo, F. and Ayadi, M., Digital flatness-based robust controller applied to a thermal process, *IEEE International Conference on Control applications*, Mexico, pp.936–941, 2001.
- [8] BEN ABDALLAH, M., AYADI, M. et BENREJEB, M., Flatness-based control of MIMO linear Systems, *Transaction on Systems, Signals and Devices*, vol. 6, n. 1, pp.1–24, 2011.
- [9] SLEIMI, M., BEN ABDALLAH, M. and AYADI, M., Digital flatness-based control design for LTI MIMO systems, *IEEE 4th International Conference on control Engineering and Information Technology*, CEIT, Hammamet, 2016.
- [10] BENABDALLAH, M., Sur la commande par platitude de systèmes dynamiques SISO et MIMO, *Thèse de Doctorat en Génie Electrique*, Ecole Nationale d'Ingénieur de Tunis, 2014.
- [11] Fliess, M., Levine, J., Martin, P. and Rouchon, P., Flatness and defect of non-linear systems: introductory and examples, *International Journal of Control*, vol. 61, no. 6, pp.1327–1361, 1995.
- [12] Kamen, E. W., Fundamentals of linear time-varying systems, *The control handbook, Control System Advanced Methods*, second edition, Taylor and Francis Group, New York.
- [13] Malrait, F., Martin, Ph. and Rouchon, P., Dynamic feedback transformations of controllable linear time-varying systems, *nonlinear Control in the year 2000. Edite par Isodori, A., Lamnabhi- Lagarrigue, F., Respondek, W., Lecture Notes in Control and Information Sciences*, vol. 259, pp. 55–62, Springer, London, 2001.
- [14] BINGULAC, S. and MEADAWES, H.E., Controllability and observability in time-variable linear systems, *SIMA Journal of control and optimization*, no. 5, pp.64–73, 1967.
- [15] SUN, H., BUTT, S.S. and ASHERMAN, H., Discrete-time flatness-based control for a twin rotor helicopter with an extended kalman filter, *IEEE International Conference on Advanced Intelligent Mechatronics*, Banff, AB, Canada, 2016.
- [16] PAULO, S. and ROUCHON, P., Flatnessbased control of a single qubit gate, *IEEE transactions on Automatic control*, vol. 53, pp.775–779, 2008.
- [17] HUANG, R., Output feedback tracking control of nonlinear time-varying systems by trajectory linearisation, *Phd thesis*, the Russ College of Engineering and Technology of Ohio University, 1967.
- [18] FIACCHINI, M. and MILLERIOUX, G., Dead-Beat functional observers for discrete-time LPV systems with unknown inputs, *IEEE transactions on Automatic control*, vol. 58, pp.3230–3235, 2013.

Comparative Performance Analysis of Efficient MIMO Detection Approaches

Muhammad Faisal, Fazal Wahab Karam, Ali Zahir
Department of Electrical Engineering
COMSATS University Islamabad, Abbottabad Campus
Abbottabad, Pakistan

Sajid Bashir
Department of Electrical Engineering
National University of Technology Islamabad
Islamabad, Pakistan

Abstract—The promising massive level MIMO (multiple-input-multiple-output) systems based on extremely huge antenna collections have turned into a sizzling theme of wireless communication systems. This paper assesses the performance of the quasi optimal MIMO detection approach based on semi-definite programming (SDP). This study also investigates the gain obtained when using SDP detector by comparing Bit Error Rate (BER) performance with linear detectors. The near optimal Zero Forcing Maximum Likelihood (ZFML) is also implemented and the comparison is evaluated. The ZFML detector reduces exhaustive ML searching using multi-step reduced constellation (MSRC) detection technique. The detector efficiently combines linear processing with local ML search. The complexity is bounded by maintaining small search areas, while performance is maximized by relaxing this constraint and increasing the cardinality of the search space. The near optimality of SDP is analyzed through BER performance with different antenna configurations using 16-QAM signal constellation operating in a flat fading channel. Simulation results indicate that the SDP detector acquired better BER performance, in addition to a significant decrease in computational complexity using different system/antenna configurations.

Keywords—Multiple input multiple output antennas; MIMO detection approaches; performance analysis; semi-definite programming; zero forcing maximum likelihood

I. INTRODUCTION

Wireless communication technology has seen rapid developments and unprecedented growth in the fields of computing and communication technologies during the last few decades. Wireless communication provides voice, video and data services. However, emerging services in the wireless communication are demanding more efficient network channels, high-bit rate, quality of service and higher network capacity [1]. The multimedia information traffic conveyed through the global mobile networks have been gigantic [2], [3], furthermore this tendency is put to continue, the same as suggested by VNI (Cisco- visual-networking-index) estimation [4].

Moreover, as forecast in Fig. 1 it will raise almost seven times which convert toward a CAGR (Compound-Annual-Growth Rate) of 53% in the time across 2017–2020, achieving 30.6 EB per month by 2020 [4]. As shown in Fig. 2. This blazing development is primarily stimulated by the dominance of mobile phones & gadget, tablets & laptops, and the materialization of machine-to-machine (M2M) communications [5]-[7].

Consequently, these systems have moved from Single-Input Single-Output (SISO) antenna technology to more efficient MIMO antenna technology for higher data rate and spectrally efficient wireless channels, without escalating the bandwidth or transmission power of system. However, an efficient MIMO system requires significant effort for designing efficient detectors with low computational cost. It improves data rates through spatial multiplexing and Bit Error Rate (BER) performance through diversity, which uses different detecting algorithms to decode received vectors [8].

A problem encountered in the design of optimal detector is to detect original transmitted signal (information) from noisy and faded channel in digital communication systems. In any practical scenario of information exchange between the transceivers, the designing of detector poses a big challenge to meet specifications such as minimal probability errors, computational efficiency and less complexity. Unfortunately, such type of detectors is computationally complex and often left out in favor of sub-optimal detectors. However, in many cases, the performance is considerably different during analysis of suboptimal and optimal detectors. On the other hand, the computationally efficient, cost effective optimal detection makes the optimal detectors attractive in comparison to its counterpart.

The Maximum Likelihood (ML) detector gives minimum error probability [9] but it is impractical to use higher-order modulation (16-QAM) in MIMO systems due to its exhaustive search requirements. Different linear, sub-optimal and near optimal, detectors are generally discussed to reduce the ML complexity. These detectors are commonly known as Sphere Detectors that provide optimal performance with reduced com-

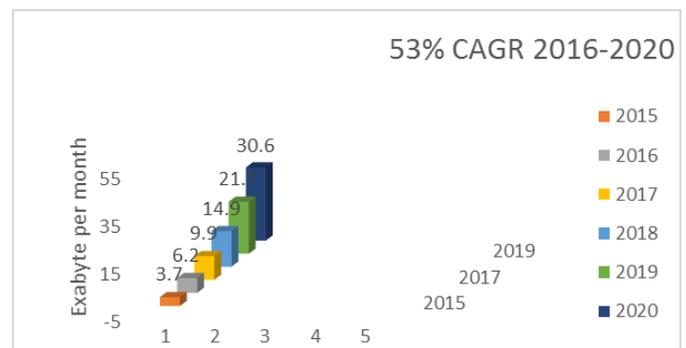


Fig. 1. Cisco VNI 2016-2020 traffic forecast for worldwide mobile data.

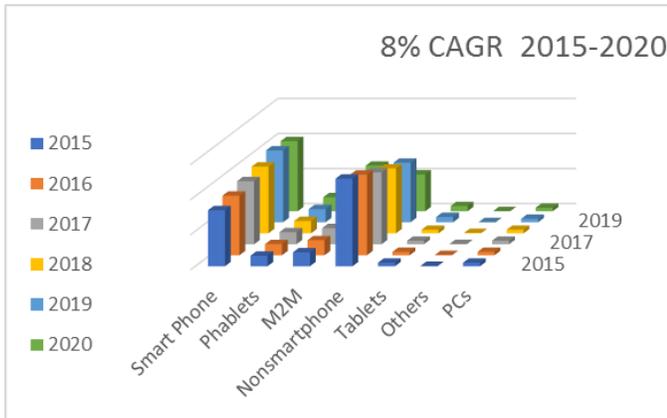


Fig. 2. Cisco VNI: Global mobile devices and connections growth forecast, 2015-2020.

putational complexity [10]. Zero Forcing (ZF) detector [11] is suboptimal linear detector. It has a polynomial complexity of cubic order $O(n^3)$ for $M_t \times N_t$ MIMO systems. It forms the computational complexity of computing the pseudo-inverse of the matrix channel. The linear, suboptimal and near optimal detectors are computationally less complex. However, the compromise is the degraded BER performance in comparison to the ML detector. Presently, computationally efficient and lesser complex high-performance MIMO detector such as ZFML [12] is near optimal heuristic detector. It reduces exhaustive ML searching using MSRC detection technique. However, the performance of ZFML detector is better in large search space.

The designing of detection poses a big challenge to meet specifications such as minimal probability errors, computationally efficient and reduce complexity. The use of semi-definite relaxation (SDR) offers efficient, high-performance detection approach [13]. SDR is efficient in solving the computationally complex ML detection problem and numerous detection problems that are discussed in [14].

The work [15] formulated ML problem in a higher dimension to relax rank-1 constraint (non-convex problem to convex problem) and derived as SDR. It provides better solution in computational complex problems [16].

II. MULTIPLE INPUT MULTIPLE OUTPUT SYSTEM MODEL

MIMO system gives significant improvement in spectral efficiency of wireless channel without escalating the bandwidth or transmission power of system. However, an efficient MIMO system requires significant effort for designing efficient detectors with low computational cost.

The block diagram in Fig. 3 depicts complex MIMO system [17]. The data bits are encoded and interleaved for transmission. The data symbols (QAM symbols) are mapped through interleave code-words; space-time encoder seizes data symbols and generates spatial data streams. The Space-Time Block-Encoder map the spatial streams and then transmit information signal to the receiver, subsequently received vector is decoded, de-mapped and de-interleaved.

The narrowband MIMO channel consists of point-to-point wireless system of M_t “transmit(T_x)” and N_r .

“receive(R_x)” ($N_r > M_t$) as shown in Fig. 4 where M_t input symbols $S_t = [S_1 \dots S_M]$, ($n_1 \times 1$), transmitted during the j^{th} time slot. N_r is having receive vectors $y_r = [y_1 \dots y_r]$, ($n_1 \times 1$). Noise is denoted as $n = [n_1 \dots n_r]$, ($n_1 \times 1$) containing AWGN elements with $\sigma^2 n$ variance. $H = [h_1 \dots h_N]$ denotes complex $M_t \times N_r$, channel matrix, the i^{th} column and i^{th} row is h_{ij} . Complex Gaussian is Rayleigh distribution that depicts flat fading channel.

MIMO techniques as shown in Fig. 6 are used in technologies e.g. Wi-Fi and LTE, and emerging techniques e.g. LTE Advanced. Comparing the performance while applying multiple-input-multiple-output (MIMO) techniques particularly, several setups with various MIMO algorithms are considered.

A. MIMO Detection Approaches and Challenges

As indicated by Shannon “Primary dilemma in communication is to replicate at some point both precisely & roughly a signal chosen on a different spot” [18].

Comparatively to typical single-input and single-output (SISO) systems, MIMO systems contain multiple interfering symbols/messages conveyed simultaneously, furthermore, subsequently these messages/symbols are anticipated to be decoded /detected at the receiver pertaining to corruption by haphazard interference/noise as presented in Fig. 5. The compound messages/symbols might be sensed/ detected alone otherwise mutually. Contrasting to alone sensing/detection, every symbol/message has to be sensed/detected considering the uniqueness of the other messages/symbols in mutual sensing/detection. Since a useful outcome, characteristically mutual detection is able of achieving a considerably superior efficiency than alone detection/sensing, even though mutual sensing/detection inflicts higher computational complexity.

The mutual sensing/detection of compound messages in Multiple Input Multiple Output systems is of fundamental significance for the purpose of grasping the important benefits of diverse Multiple Input Multiple Output methods. Due to the CCI (co-channel-interference) usually encounter in Multiple Input Multiple Output systems make up primary restrictive feature [19]-[21]. Desolately, the best possible MIMO detection issue is established as non-deterministic polynomial-time hard (NP-hard) [22], [23]. Consequently all well-known algorithms considered for resolving the problem for optimal solutions,



Fig. 3. Block diagram of Complex MIMO system.

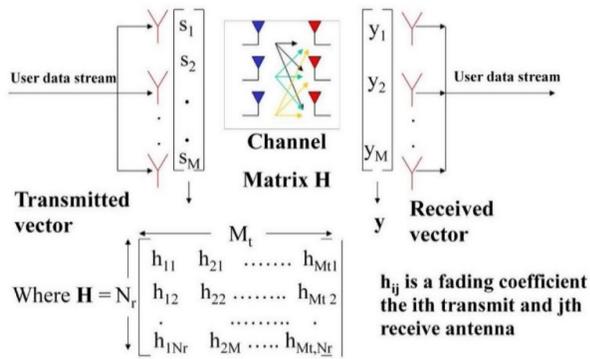


Fig. 4. MIMO channel “where M_t represents number of “transmit (T_x)” and N_r “receive (R_x)” antenna correspondingly”.

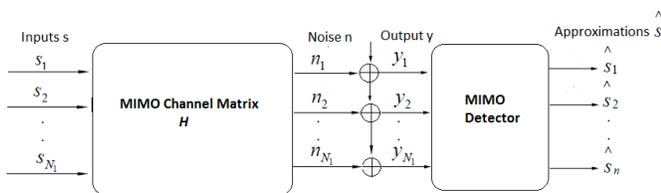


Fig. 5. MIMO detection dilemma.

incorporate exponential rise of complexity with the raise in number of decision factors. As a result, the computational complexity of the best possible ML (maximum-likelihood) condition based MIMO detection algorithms rapidly turn into redundant as the numbers of decision variables are augmented. Practically all modern ICs meet an integration density margin due to the maximum bearable internal temperature forced by the extreme power consumption, resulting restriction on IC development. As a result, one may not merely depends on Moores law. Moreover yet mild complex Multiple Input Multiple Output sensing/detection methods might be excessively power starving designed for systems based on battery. Therefore modest complexity, however superior-performance sub-optimum Multiple Input Multiple Output detection procedures are required intended for realistic Multiple Input Multiple Output-applications.

Spatial multiplexing methods concurrently transmit self-determining information sequences, frequently known layers, using multiple antennas. With an M transmit antennas, in general bit rate contrast to a single-antenna system is improved by a factor of M with no requirement for additional bandwidth or further transmission power. Channel coding is frequently engaged, hence to warranty a definite error performance. As the individual layers are super-imposed throughout communication, need to be alienated at the receiver by an interference cancellation category of algorithm (classically in combination with multiple receive antennas). A renowned spatial multiplexing method is the BLAST (Bell-Labs Layered Space-Time Architecture). The realized bit rate comparing to a single-antenna system is known multiplexing gain e.g. an antenna gain, multiplexing gain and diversity gain.

B. The multiplexing gain

The truth to facilitate the capacity of a MIMO system with M transmit and N receive antennas raises (more or less) linearly with the minimum of M and N (exclusive of entailing further bandwidth or additional transmission power) is an fascinating outcome. For SISO, setting a predetermined bandwidth, capacity may barely be improved logarithmically with the SNR, by rising the transmit power. In [1], the theoretical capacity outcomes for MIMO systems were matched by the scheme of the BLAST method, achieving bit rates approximately 90% of outage capacity. The first real-time BLAST demonstrator was set with $M = 8$ transmit and $N = 12$ receive antennas achieving exceptional bit rates of 40 bit/s per Hertz in contrast to any SISO system.

C. Spatial Diversity

Multiple antennas may also be used to enhance the error rate of a system (error performance), as a result of transmitting and receiving unneeded signals presenting the identical information sequence accommodating in the spatial domain, rather than in the time domain without lowering the effective bit rate in contrast to single-antenna transmission. Spatial diversity methods primarily aim at an enhanced error performance in contrast to spatial multiplexing schemes based on a diversity gain and a coding gain. Two forms of spatial diversity as macroscopic and microscopic diversity can be found in a comprehensive survey of spatial diversity for wireless communication systems [20].

D. Signal-to-noise Ration and Co-Channel Interference

In addition to higher bit rates and smaller error rates, multiple-antenna techniques may also be employed to enhance the SNR at the receiver and to contain co-channel interference in a multiuser situation by adopting smart antennas or software antennas. Beam-forming schemes which are interpreted as linear filtering in the spatial domain are employed, the beam patterns of the transmit and receive antenna array may be steered in particular preferred directions, whereas un-preferred directions having significant interference may be nulled.

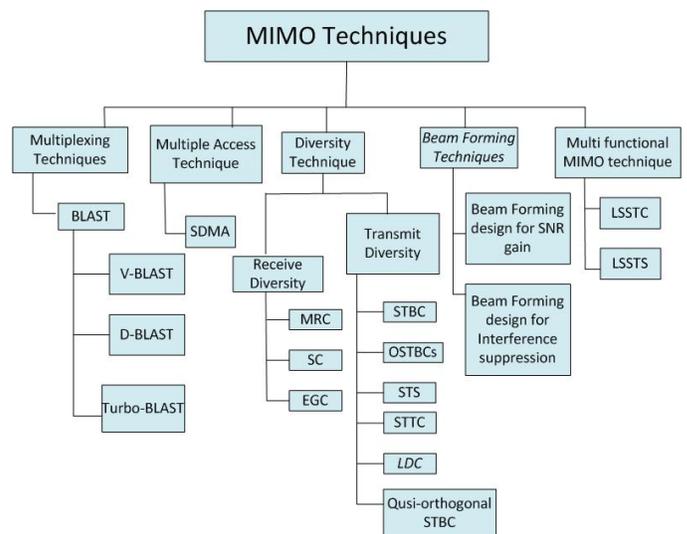


Fig. 6. MIMO Techniques.

E. Smart antennas and beamforming Schemes

Beam-forming schemes may be used to achieve enhanced data rates and better error rates causing better SNR at the receiver and suppressed co-channel interference (CCI) in a multiuser scenario. As the spectrum is restricted, so the sharing is needed to enhance the capacity of cell by allocating the available bandwidth at the same time to multiple users using multiple access methods while maintaining the quality of service trade within the existing users. SDMA employed the spatial separation of the mobile users to enhance the use of the frequency spectrum. The transmission power of every user is restricted by Space division multiple access using spot beam antenna serving by the similar frequency or diverse frequencies. TDMA or CDMA is employed by the antenna beam covering diverse areas, frequency may be re-used, in, for diverse frequencies FDMA may be employed. Multi-functional MIMO mingle the benefits of numerous MIMO methods, e.g. multiplexing gains, diversity gains, and beam-forming gains, e.g. V-BLAST is able of realizing maximum possible multiplexing gain, while STBC may realize the full possible antenna diversity. In V-BLAST and STBC is combined to offer together antenna diversity and spectral efficiency gain. Additionally, combined array processing is enhanced through enhancing the decoding order of the different antenna layers. The optimized receive diversity gain for the mutual V-BLAST- STBC system assisted by the number of separately fading diversity channels is achieved in an iterative decoding algorithm. A transmission scheme known as D-STTD (double space-time transmit diversity), comprising of two STBC layers at the transmit antennas, whereas the receiver is prepared with two antennas is presented. Beam-forming has pooled both spatial diversity and spatial multiplexing schemes to realize extra performance gain e.g. beam-forming and STBC has been pooled together.

III. MULTIPLE INPUT MULTIPLE OUTPUT SENSING/DETECTION DILEMMA

Regardless of the reality that related problems have been identified for a while [22], [24], [25] the idiom “MIMO detection” became common primarily with the advent of multiple-antenna systems throughout the mid-1990s [26], [27]. As a consequence, in the common sense, Multiple Input Multiple Output detection typically applied on to the symbol detection issue materialized in narrow-band SDM based multiple antenna methods, e.g. VBLAST (Vertical Bell Laboratories layered space-time system) [28]. Though, it is emphasized to facilitate a group of significant signal processing methods, Multiple Input Multiple Output detection ought to be understood based on a general mathematical model. In the broad sense, the MIMO detection issue can be characterized for an M_t input linear scheme whose transfer function is expressed by a matrix, containing non-orthogonal columns and Its output N_r is corrupted by additive random interference, which does not essentially comply with the Gaussian distribution. The compound inputs may be represented like a vector S that is arbitrarily retrieved from the set S^{M_t} complied by M_t element vectors, whose elements are drawn from a finite set $S_t = [S_1 \dots S_M]$. The “deductive” or “presumptive” probability of picking each vector from S^{M_t} is similar. The set is generally referred as the constellation alphabet, whose components may get any complex or real values. Furthermore, $S_n, n = [n_1 \dots n_r]$,

correspond to the realizations of S , therefore these are the components of S^{M_t} . Next the relationship among the inputs and the outputs of this linear scheme may be described by

$$Y = Hs + n \quad (1)$$

where Y is receive signals vector, H is channel matrix of the system, and n additive noise is denoted by $\sigma^2 n$ containing AWGN elements with variance. Based on the particular applications it may be moreover the field of \mathbb{R} (real numbers) or the field of \mathbb{C} (complex numbers). In brief, every scheme having compound inputs & outputs, and pertaining to additive random interference may be considered as a Multiple Input Multiple Output system, however the MIMO detection issue concerned in MIMO systems, is simply just tackled whose channel matrix is non-orthogonal in columns. This is significant that the s (constellation alphabet), the M_t (number of inputs) and the number of outputs N_r are usually considered as constant quantities for a particular system. Therefore, these are understood to be identified by default, though it will not be explicitly underlined, except needed. While an additional note, as the input message/symbol vectors of compound successive time-slots are linked together using space-time-coding [29], [30], the Multiple Input Multiple Output system is specified as

$$Y = Hs + n \quad (2)$$

wherever Y denoting a matrix indicating the message/signal received in multiple time-slots, H denoting a matrix denoting the space-time codeword, and presents the resultant noise/interference matrix. Equation (1) may be realized from (2) by putting the number of time-slots regarded to one. In this context, (2) is more general than one (1), though, equation (2) is primarily employed for differentiating space-time-coding aided Multiple Input Multiple Output schemes. This is due to the best possible ML decoding/detecting could be purely implemented based on the separate symbol-by-symbol decoding approach or pair-wise decoding approach [30], [31]. Thus, majority cases related with Multiple Input Multiple Output detection, depends on the system model presented in (1). The fundamental job of Multiple Input Multiple Output detection is to approximate the key input vector based on the information of the expected/received signal vector Y and the channel matrix H . If the instantaneous value of H is eminent from precise channel estimation, the denote/detection of s is said to be based on coherent detection. Even though, if the precise estimation of the instantaneous channel state is evaded, the detection of s fit in to non-coherent detection scheme.

IV. MULTIPLE INPUT MULTIPLE OUTPUT DETECTION APPROACHES

A. Maximum Likelihood detection

The ML detection in higher-order modulation (M-QAM) is an NP-Hard problem due to exhaustive search in MIMO systems. Therefore, it is impractical even for moderate systems. For this reason, less computationally complex and efficient detectors are needed to develop.

The transmit symbols are from a random finite alphabet or constellation $S \subset \mathbb{C}$, $S = S_1 \dots S_M$ of size M_t . The detector's function is to select one of the M^{M_t} or 2^{kM} possible

transmitted symbol vectors from whole set of transmitted symbols. Suppose symbol vectors $S \in S^{M_t}$ are equi-probable.

$$S^* = \underset{s \in s^{n_t}}{\arg \max} P(y \text{ is observed} | s \text{ was sent}) \quad (3)$$

ML detector always returns an optimal solution according to (3). Optimal detection is performed over the search space of all possible input s vectors. Since the search space has random integer components, this problem called least-squares optimization problem and it is non-deterministic polynomial (NP) which is time hard and Combinatorial Optimization Problems (COP). This type of problems involves an optimal solution with respect to an objective function for detection. COP use exhaustive search to enumerate optimal solutions and selecting the one which minimizes the objective function in shown in (4).

$$S^* = \underset{s \in s^{n_t}}{\arg \min} P \|Y - Hs\|^2 \quad (4)$$

The ML detector of (4) represents a discrete optimization problem over $|S|^{M_t}$ candidate vectors $S \in |S|^{M_t}$. Unfortunately, such problems are in general hard to solve and it has been shown that the problem of (4) for general y and H is NP-hard [23].

ML detector for $M_t \times N_r$ MIMO system with higher-order modulation (M-QAM constellation alphabet), has high computational complexity that increases exponentially with constellation size M and number of transmitters M_t . A ML detector has to search $|M|^{M_t}$ symbols vectors. The ML computational complexity in 16-QAM and 2 transmit antennas is $|M|^2 = 16^2 = 256$, for 3 transmit antennas, it is $|M|^3 = 16^3 = 4096$ and for 4 transmit antennas, it is $|M|^4 = 16^4 = 65536$.

B. Zero Forcing detection

ZF is a suboptimal linear detector which uses Moore-Penrose pseudo-inverse i.e $H^\dagger = (HH^H)^{-1}H^H$. Here, HH^H is the channel frequency response of the received signals, perfectly suppressing the Inter-Symbol Interference (ISI). For example, frequency response $F(f)$ of the detector $s(s)$ is constructed as $s(s) = 1/F(f)$. Thus, the combination of channel and equalizer gives a linear phase $F(f)s(f) = 1$, meaning a flat frequency and channel response i.e., $H(s)$. Afterwards, the input signal is multiplied by the reciprocal of this. This removes the effect of ISI from the received signal.

ZF is successive technique to cancel interference or ISI. The interference caused by transmitted channel is then subtracted from the received signals(s). For simplicity let us consider MIMO channel modeled as in (1). To get input symbols (s), we need matrix that satisfies $H^\dagger H = 1$. The ZF detector to meet that type of constraint is given by,

$$H^\dagger = (H^H H)^{-1} H^H \quad (5)$$

Where H^\dagger is Equalization Matrix and H is Channel Matrix. Equation (5) is known as the Pseudo-inverse of $M_t \times N_r$

matrix. Here,

$$\begin{aligned} H^H H &= \begin{bmatrix} h_{1,1}^* & h_{1,2}^* \\ h_{2,1}^* & h_{2,2}^* \end{bmatrix} \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} |h_{1,1}|^2 + |h_{2,1}|^2 & h_{1,1}^* h_{1,2} + h_{2,1}^* h_{2,2} \\ h_{1,1}^* h_{1,2} + h_{2,1}^* h_{2,2} & |h_{1,2}|^2 + |h_{2,2}|^2 \end{bmatrix} \end{aligned} \quad (6)$$

To observe the matrix, $H^H H$ are not zero in off diagonal elements because the off-diagonal elements are non-zero in values. ZF detector try to null out the interfering terms when performing the detection, i.e. when solving for s_1 , the interference from s_2 is tried to make it null and vice versa. ZF performs $N_r - M_t + 1$ diversity order in a $M_t \times N_r$ MIMO system with N_r possible diversity order. The ZF degrades BER performance due to noise amplification, lost whiteness property of AWGN, correlated across the data streams and is unable to detect parallel received signal. ZF detector [11] is suboptimal linear detector. It has a polynomial complexity of cubic order $O(n^3)$ for $n \times n$ MIMO systems. It forms the computational complexity by computing the pseudo-inverse of the matrix channel H .

C. Near Optimal Heuristic Approach

The quadratic form of (3) given as

$$f(x) = \|y - Hs\|^2 \quad (7)$$

The function $f(x)$ in (5) is convex. This near optimal Heuristic detector or algorithm reduces exhaustive ML searching and is suitable for higher order constellations. This detection algorithm also termed as multi-step reduced constellation (MSRC) detection performs local search of the target symbols within certain constraint specified reduced search space. In fact, a ZF initial solution estimate is used to define the radius of search. Constellation points around the ZF solution are searched in steps using (3) to find out the minimum Euclidian distance. This particular method which starts with the ZF processing is termed as ZFML [12].

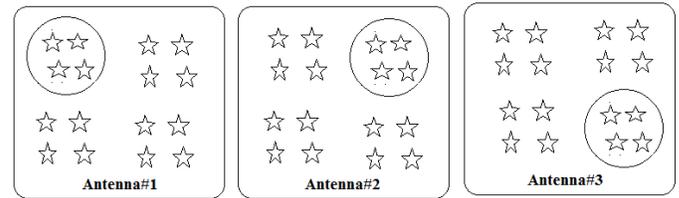


Fig. 7. ZFML reduced constellation search.

First y is computed and then a ML search around the neighborhood of y is performed as depict in Fig. 7. Each of the M_t symbol generates a neighbor list, then a joint ML search our reduced constellations is performed. This process continues in an iterative fashion. Since the search output is generated until optimal solution is achieved. In contrast to ML search over the entire search space, the ZFML uses reduced constellation to decrease the computational complexity factor $(M/M_n)^{M_t}$ of ZFML, where M constellation size is, M_n is reduction rate and is number of transmitting antennas. If reduction rate is $|M|^{M_t} = 16$, then after completing two rounds, the reduction rate is $16^{M_t}/2$.

D. Sub-Optimal Approach

Comparing to other MIMO detectors, the SDR technique is based on a respite of the best possible Multiple Input Multiple Output detection issue to the mathematical model of semi-definite programming (SDP), which is a sub-field of convex optimization [33].

Convex optimization represents a subfield of the general mathematical optimization issue. This provides study of fundamental framework as shown in Fig. 8 for minimization of a convex objective function over convex sets. It solves mathematical optimization problems by means of convex optimization which is considered as straightforward problem, due to powerful numerical algorithms, e.g. the interior-point method [35], which efficiently compute the optimal solution of convex issues. Thus, convex optimization problem is resourcefully resolved, in contrast to non-convex optimization problem which is usually tricky to work out. Convex optimization has a variety of additional vital characteristics e.g. each local optimal resolution represents the global most favorable result; consequently, there is no hazard of being misled by the local best possible. Furthermore, a thorough optimal situation and the duality theory is present to substantiate the best nature of a resolution in convex optimization problem [34], [36].

The SDR based MIMO detectors in recent times have received considerable research interest [37], [38]. The main striking feature of the SDR detectors is to support a polynomial-time 16 worst case computational complexity, whilst attaining a soaring performance in certain situations. SDR was first suggested for a BPSK modulated CDMA scheme [37], [39], [40], moreover next it was extended to Quadrature phase shift keying (QPSK) [41].

SDR [15] is suboptimal detection technique for higher order modulation (M-QAM) in MIMO system. SDR is efficient in solving the computationally complex ML detection problem and numerous detection problems are discussed in [42]. However, [15] formulated ML problem in a higher dimension and afterward relax rank-1 constraint (non-convex problem to convex problem) and derived as SDR, the rank relaxation method is known semi-definite programming (SDP). The SDP is better in computational complex problem and solve the problem efficiency in polynomial time [43]. The fundamental principle of SDP based detectors is demonstrated in Fig. 9, where the boxes signifies the technical challenges.

1) Rank Relaxation: There are several engineering problems having non-convex constraint such as NP-hard problems. In such problems, the non-convex constraints may be dropped or relaxed, resulting in a relaxed problem i.e. convex. Drop-

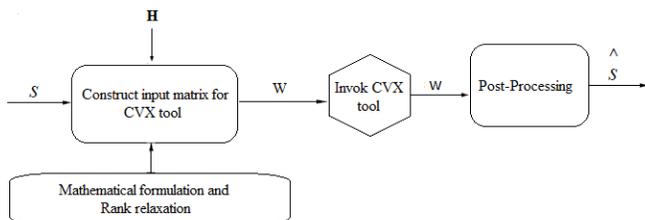


Fig. 8. Framework of solving problem using convex optimization.

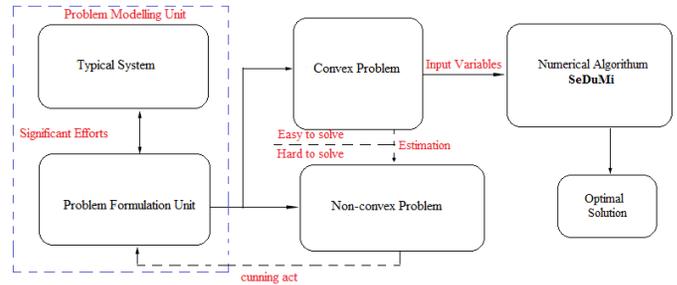


Fig. 9. The basic principle of the SDR detection technique.

ping constraints generate more feasible set to solve problem (minimize or maximize the objective function). Obviously, this feasible set gives more solutions which are not desirable in optimal solution. The reason is the relaxation of the constraints, and the set cannot be directly used as an estimated solution of the original problem, because it may not lie in the original feasible set. Thus, simple quantization, Eigen-value decomposition and randomization are used to approximate the solution [15]. Therefore, a relaxation algorithm solves the relaxed problem.

2) Semi-definite Programming Rank Relaxation: Afterwards, an approximation algorithm is used to transform a relaxed solution to an approximate one for the original problem. Considering ML problem as optimization problem in SDP-Relaxation, i.e.

$$\hat{s}_{ML} = \arg \min_{\hat{s} \in s^{ML}} \|y - Hs\|^2 \quad (8)$$

The SDR attempts to estimate the solution of (8) by forming a non-convex to convex problem. The problem (8) can be expressed as:

$$\|y - H\hat{s}\|^2 = \hat{s}^T H^T \hat{s}^T - 2y^T H \hat{s} + y^T y \quad (9)$$

$$\begin{bmatrix} A & b \\ b^T & c \end{bmatrix} \quad (10)$$

Lemma: 1 Lemma: 1 let A be a symmetric matrix. The condition $s^T A s - 2b^T s + c \geq 0$ holds for all s if and only if the matrix $\begin{bmatrix} A & b \\ b^T & c \end{bmatrix} \geq 0$ is semi-definite positive.

The problem (9) can be written as quadratic function using lemma#1

$$\begin{bmatrix} s^T & 0 & 1 \end{bmatrix} \begin{bmatrix} H^T H & 0 & -H^T y \\ 0 & 0 & 0 \\ -y^T H & 0 & y^T y \end{bmatrix} \begin{bmatrix} s \\ 0 \\ 1 \end{bmatrix} \quad (11)$$

Consider $w = [s \ 0 \ 1]^T$ to form the ML problem as optimization problem

$$\min_{w \in \mathbb{R}^{m+1}} w^T C w \quad (12)$$

where

$$C = \begin{bmatrix} H^T H & 0 & -H^T y \\ 0 & 0 & 0 \\ -y^T H & 0 & y^T y \end{bmatrix} \quad (13)$$

$s = \pm 1, \pm 3$. In symmetrical problem constraints that given by $[w]_{m+1} = 1$ do not need to be enforced due to the symmetry of the problem such that if \tilde{x} is a minimizer of (12) then so is $-\tilde{x}$. Therefore $w^2 = 1$ implies $w \in \{\pm 1\}$. By introducing $W = wTw$ the problem can be equivalently written as optimization problem.

$$\begin{aligned} & \min_{W, x} CW \\ & \text{s.t. } \text{diag}(W) = e \\ & \quad W = ww^T \end{aligned} \quad (14)$$

Problem (8) and (14) are equal in solution.

The problem (14) also written as with the rank of W

$$\begin{aligned} & \min_W CW \\ & \text{s.t. } \text{diag}(W) = e \\ & \quad \text{rank}(W) = 1 \end{aligned} \quad (15)$$

since vector e is all one which satisfy $e = \text{diag}(W)$ factorized according to $W = ww^T$ for $w \in \{\pm 1\}^{m+1} = S^{m+1}$.

Problem (8) and (15) are equal. Specifically, these problems are computationally complex and NP-hard [23] to solve. The problem (14) is non-convex problem due to rank-1 constraint. In [16] relax rank-1 constraint from W with a, positive semi-definite, constraint, $W \geq 0$. Now the semi-definite relaxation of (15) written as

$$\begin{aligned} & \min_W CW \\ & \text{s.t. } \text{diag}(W) = e \\ & \quad W \geq 0 \end{aligned} \quad (16)$$

However, $W = ww^T$ implies $W \geq 0$ it follows the problem (15) shows a relaxation of (14). While W have $(m+1)^2$ variable, as different to $m+1$ variables in w , the relaxation of rank-1 takes place in higher order dimensional.

The semi-definite Problem (16) has efficient methods to solve in polynomial time [46]. In particularly there are efficient techniques outlined in [15], [32], [46], [47] which solve (16) in $O(K^2N)$ time. If the solution of problem (16) occurred to be rank-1 then it is solved in (15). On the basis of studying the problem (16) in context of digital communication, sometimes it gives same solution of (16) is certainly of rank-1. In a scenario, when this is not the case of the solution to (15), even then this can be guaranteed to obtain from the solution of (16) having high probability. However, an efficient method for the estimation of \hat{s}_{ML} is proposed for high rank solution in [42], [49] and the very method had analyzed for accuracy in the work [48].

3) *Optimization problem Comprised in SDP*: The problem is combinatorial problem/optimization problem with finite alphabet constraints. The problem may be solved in brute-force fashion by searching over all the $|M|^M$ possible vector combination.

$$\begin{cases} \min_{s,t} \|y - Hs\|^2 \\ \text{s.t. } s_i \in \{\pm 1, \pm 3\}, i = 1 \dots 2k. \end{cases} \quad (17)$$

The finite alphabet constraint $\pm 1, \pm 3$ can be replaced with the polynomial constrain $(s_i + 1)(s_i - 1)(s_i + 3)(s_i - 3) = 0$, $i = 1 \dots 2k$ and introduce the slack variables $t_i = s_i^2$ for, $i = 1 \dots 2k$ to formulate in a higher order polynomial constraint.

$$\begin{cases} \min_{s,t} \|y - Hs\|^2 \\ \text{s.t. } \dots s_i^2 - t_i = 0, i = 1 \dots 2k \\ \dots t_i^2 - 10t_i + 9 = 0, i = 1 \dots 2k \end{cases} \quad (18)$$

SDR detector approximates expression (18) by relaxing the feasible set of expression (17) and forming a non-convex to convex problem for optimal solution. To formulate higher dimension optimization problem, it derives the SDR and replaces s and t vectors of expression (18) with rank-1 semi-definite matrix $W = ww^T$, where $w^T = [s^t t^T]$.

The constraint easily identify $W = ww^T$, $W_{2,2} = tt^T$ and $W_{2,3} = t$, where $W_{i,j}$ for $i, j = 1, 2, 3$ are the $(i, j)^{th}$ sub-blocks W of suitable sizes. However, in order to make problem in expression (18) an optimization problem we have:

$$\begin{aligned} \min_w Tr \left\{ \begin{bmatrix} H^T H & 0 & -H^T y \\ 0 & 0 & 0 \\ -y^T H & 0 & y^T y \end{bmatrix} \right\} & \quad (a) \\ \text{s.t. } \text{diag} \{W_{1,1}\} - W_{2,3} = 0 & \quad (b) \\ \text{diag} \{W_{2,2}\} - 10W_{2,3} + 91 = 0 & \quad (c) \\ W \geq 0 & \quad (d) \\ W_{3,3} = 1 & \quad (e) \\ \text{rank}(W) = 1 & \quad (f) \end{aligned} \quad (19)$$

The problem in expression (19) is optimization problem and non-convex problem due to rank-1 constraint. However non-convex problem in expression (19) is computationally hard to solve, so relax the constraint 1 to form convex problem:

$$\begin{aligned} \min_w Tr \left\{ \begin{bmatrix} H^T H & 0 & -H^T y \\ 0 & 0 & 0 \\ -y^T H & 0 & y^T y \end{bmatrix} \right\} & \quad (a) \\ \text{s.t. } \text{diag} \{W_{1,1}\} - W_{2,3} = 0 & \quad (b) \\ \text{diag} \{W_{2,2}\} - 10W_{2,3} + 91 = 0 & \quad (c) \\ W \geq 0 & \quad (d) \\ W_{3,3} = 1 & \quad (e) \end{aligned} \quad (20)$$

Note that the problem in expression (20) leads to a linear objective. This is subjected to use of equalities and inequality of a linear matrix. This type of SDP problem can be solved using CVX tool in polynomial time [16].

4) *Complexity analysis*: The majority common techniques for resolving SDP issues of modest sizes are IPMs e.g. DSDP [51], SeDuMi [52], SDPA [53], etc. whose computational complexities are polynomial. Semi-definite programs of realistic size may be resolved in polynomial time within any precise precision by IPMs which are iterative algorithms using Newton-like techniques to produce search directions for finding an estimated resolution to the nonlinear system. As the IPMs converge vary fast and precised best solution is attained within a polynomial number of iterations.

TABLE I. PERFORMANCE AND COMPLEXITY COMPARISON OF MIMO DETECTORS

Tech	Scheme	Mult. Ant	CSI	BW	Rx Compl.	Benefits	Remarks
SM	V-BLAST [29]	Tx & Rx (N/M)	Rx	N	Moderate	MUX	Moderate complexity, Improved performance and diversity order
	D-BLAST	—	—	N	Moderate	MUX & DIV	
	Turbo-BLAST	Tx & Rx	—	N/W	Moderate	MUX & DIV	
SD	MRC	Rx	Rx	N	Low	DIV	Low complexity and Improved performance and diversity order
	OSTBCs	Tx (Rx optional)	—	N	Low	DIV	
	STBCs	—	—	N	Low	DIV	
	Linear dispersion codes	—	—	N	Moderate	DIV and/or MUX	Moderate complexity,
	ST-IDM	—	—	N/W	Moderate	DIV	Improved performance.
	TR-STBC	—	—	W	Moderate/high	DIV	High complexity, Improved performance and diversity order.
	STTCs	—	—	N	Moderate/high	DIV & COD	
	Delay diversity	—	—	N	Moderate/high	DIV	
	STTCs	—	—	N	high	DIV & COD	
	Differential ST schemes	—	—	N	Varies	DIV (& COD)	Quasi-ML performance and diversity order,
SF/ STF codes	—	—	No CSI	W	Varies	DIV (& COD)	Low average complexity, High complexity(worst-case),
SA	Rx beamforming	Rx	Rx	N/W	Low	ANT	Low complexity and Improved performance and diversity order.
	Tx beamforming	Tx	Tx &	N/W	Low	ANT	
	Limited feedback schemes [54]	Tx (& Rx)	RxTx (lim.) & Rx	N	Varies	ANT & DIV/MUX	High complexity(worst-case)

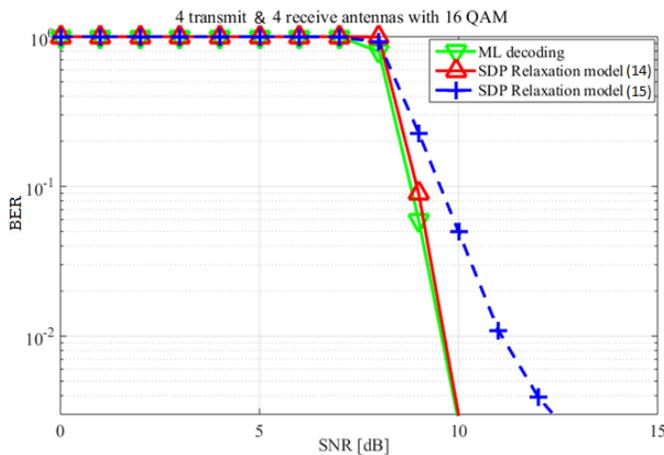


Fig. 10. Complexity comparison.

In our numerical analysis, addition of the non-negativity limitation raises the computational complexity of the using DSDP and SDPA for solving (14), and SeDuMi is implemented for solving (15). Analysis for the worst case complexity of solving models (15) and (16) by IPMs is presented here. The SDP model is devised as a typical linear cone program using slack variables addition for solving relaxation (16) and the linear conic issue using the optimization software SeDuMi [50]. The extra inequality constraints construct the model in (15) significantly sturdy than the model in (15), however too further hard to resolve. The issue in (16) is tractable as the problem sizes of our concern are modest considering a trade-off among the strength of the bounds and the computational. The randomization process performed here to reinforce the bound achieved is insignificant. Though, utilizing the structure and sparsity feature of semi-definite programs may be vital to the proficient computation of their solution. Every constraint matrices in relaxation models (15) and (16) are rank-one reducing the complexity of interior point algorithms for positive semidefinite programming converging linearly resulting reduction in computation time and memory needs. Fig. 10 depict the performance of (15) and (16).

Various MIMO detectors have different performance-and-complexity profiles having pros and cons. It seems to be a good time now, after reviewing the state of the art, to establish

some comparison amongst all these methods as shown in Table I. We studied a qualitative comparison of the performance and complexity features of the MIMO detectors, and then reviewed their analytical performance and complexity results We then, extended that table (Table I), which depicts the whole picture, listing the proposals strength points and eventual drawbacks. As we can see from Table I, not all methods have the same level of technological consolidation, particularly in terms of signaling and essential necessities.

V. SIMULATION ENVIRONMENT AND PERFORMANCE ANALYSIS

The software tool CVX is used for the performance analysis of the optimal MIMO detection approaches based on SDP. It is modeling tool built on top of MATLAB. It is powerful tool for modeling a prototype and algorithms incorporating convex problems using DCP method [45].

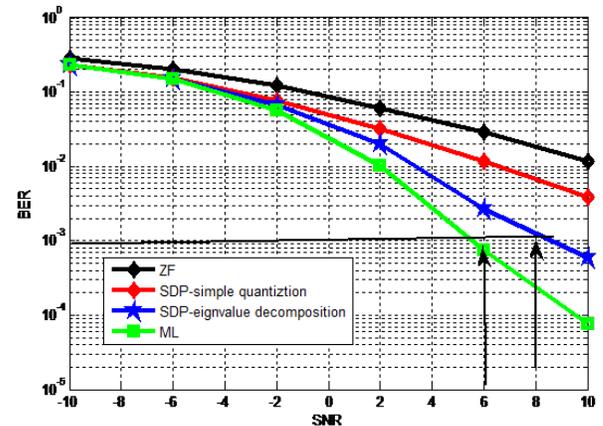


Fig. 11. BER Performance of selected MIMO hard detectors in a 3×3 systems using 16-QAM signal constellation in a flat-fading channel.

Extensive simulations were run to evaluate and analyze the performance of the quasi optimal MIMO detection approach based on SDP. The results in Fig. 11 depict BER versus SNR for the 3×3 MIMO system using 16-QAM ($16^3 \approx 4096$) in a flat fading channel. For comparison, we have simulated the optimal and linear detectors. However, all the detectors have same performance from -10dB to -4dB due to high noise effect;

it is observed that as SNR increases, noise effect decreases and BER performance improves linearly. The optimal detector ML achieves a BER of 10^{-2} at an SNR of 2dB and BER of 10^{-3} at an SNR of 6dB, by examining all possible transmitted vectors. Suboptimal approach i.e. SDP-simple quantization in a complex system achieves BER of 10^{-2} at an SNR of 6dB and SDP-Eigen-value decomposition in a complex system achieves BER of 10^{-3} at an SNR of 8dB. ZF achieves BER of 10^{-2} at a SNR 10dB. Here, its BER degraded due to noise amplification. It is easy to see the advantage of the SDR detectors over the ML and ZF detectors. However, SDP-Eigen-value decomposition approximation technique achieved considerable reduction in complexity at the cost of only 2dB.

The performance of optimal detector in term of probability is good as compared to simple detector. In [15], the author discussed the performance of different detectors exist in CDMA and SDR detector for higher order QAM constellation in [9] in the context of the MIMO channel. For complexity, simple and closed-form expressions are infrequent. The more convenient way is to model it with increasing complexity with m . The detectors complexity specified through cubic $C(m)$ and is supposed to be in $O(f(m))$, for few function $f(m)$, and $C(m) \leq cf(m)$ for $m \geq M$ where c and M are constants [44]. The complexity of a detector in $O(p(m))$ for few polynomial $p(m)$ the detector is assumed to be polynomial complex one. The complexity $C(m)$ rates are bounded by $O(\cdot)$. Still, polynomial complexity detectors are normally considered to be efficient, and the complexity of most polynomial detector is relatively small in practice [44]. Additionally, the complexity measures are obtained through simulations to analyze system performance or design.

The complexity of generic SDP approach for non-convex problem (11a)-(11f) is cubic $O(N^6.5)$ [9]. If relax rank-1 constraints as in (11f) just relax such as (12a)-(12e) and solve diagonal elements in W , then non convex problem is converted into convex problem. Hence the complexity of SDP approach for convexified problem is roughly cubic $O(N^3.5)$, where $N = 2M + 1$, while M is the number of QAM symbols[9].

The simulation results in Fig. 12 depicts BER versus SNR of SDP-simple quantization approximation technique in different system configurations using 16-QAM constellation in a flat fading channel. The performance of SDP-simple quantization in computationally complex system improves with increasing in number of transmit and received vectors. However, suboptimal approach (4×4) SDP achieved a BER of 10^{-2} at an SNR of 5dB, (3×3) SDP achieved a BER of 10^{-2} at an SNR of 6dB and (2×2) SDP achieved a BER of 10^{-2} at an SNR of 9dB.

The results in Fig. 13 depict BER versus SNR for the different antenna configuration using 16-QAM ($16^3 \approx 4096$) in a flat fading channel. However, SDP-simple quantization approximation technique in (2×4) MIMO system achieved BER of 10^{-3} at an SNR of 2dB; in (3×4) MIMO system achieved a BER of 10^{-3} at an SNR of 6dB; and in (4×4) MIMO system achieves a BER of 10^{-2} at an SNR of 2dB. However, SDP-simple quantization in (2×4) and (3×4) MIMO systems performance is improved.

The results in Fig. 14 depict BER versus SNR for the 3×3 system using 16-QAM ($16^3 \approx 4096$) in flat fading channel. For

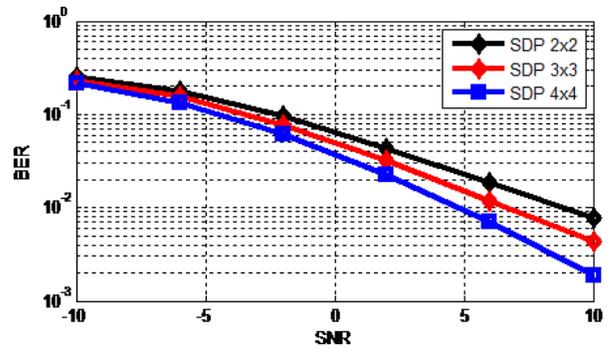


Fig. 12. BER Performance of different system configurations MIMO systems using 16-QAM signal constellation in a flat fading channel.

comparison, we have simulated the suboptimal detector with full complexity and near optimal detector with different search spaces. However, suboptimal and near optimal detectors have same performance from -10dB to -4dB due to high noise effect. It is observed that as SNR increases noise effect decreases and BER performance improves linearly.

However, sub-optimal detector SDP-simple quantization approximation technique achieved a BER of 10^{-2} at an SNR of 6dB in computationally complex system. ZFML with 16 vector search space at 2.5 neighbor size with $|2|^{Mt}$ reduction rate achieves a BER of 10^{-2} at an SNR of 13dB; ZFML with 64 vector search space at 4 neighbor size with $|4|^{Mt}$ reduction rate achieved a BER of 10^{-2} at an SNR of 10dB; ZFML with 125 vector search space at 5 neighbor size with $|5|^{Mt}$ reduction rate achieved a BER of 10^{-2} at an SNR of 9dB; and ZFML with 343 vector search space at 7 neighbor size with $|7|^{Mt}$ reduction rate achieved a BER of 10^{-2} at an SNR of 7dB. ZFML computational complexity is bounded by maintaining small search areas, while performance is maximized by relaxing this constraint and increasing the cardinality of the search space.

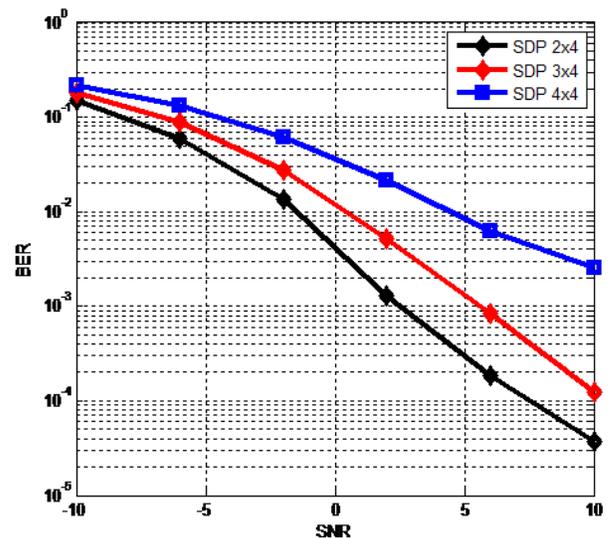


Fig. 13. BER Performance of different Antenna Configurations MIMO Systems using 16-QAM signal constellation in a flat fading channel.

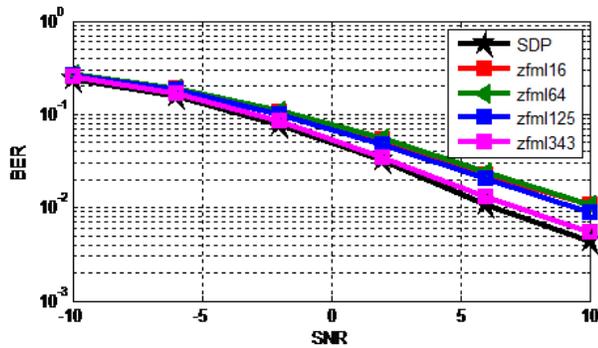


Fig. 14. BER Performance of SDP and ZFML MIMO hard Detectors in a 3×3 Systems using 16-QAM signal constellation in a flat fading channel.

However, SDP-simple quantization approximation technique achieved considerable performance in a computationally complex problem ($16^3 \approx 4096$ vectors) as compared to the ZFML detector in a less computationally complex problem $|7|^{M_t}$ ($7^3 \approx 343$ vectors). Therefore, we find the computationally efficient SDR detector as a competitive detector in comparison to other near-optimal methods.

Compared to SDP that performs a coarse search over the complete search space the ZFML used a reduced constellation, therefore its computational complexity is $(M/M_n)^M$. Where M is the constellation size, M_n is the neighbors list and M_t is the number of transmitters.

VI. CONCLUSION AND FUTURE DISCUSSION

The presented work aimed to analyze the efficiency of MIMO detection approaches both in terms of BER performance and computational complexity. Specifically, the work is focused on performance evaluation and comparison of two heuristic suboptimal detection algorithms previously proposed in literature, namely, the ZFML and the Semi-definite relaxation detectors. The presented simulation results are relating to the performance of the two algorithms including the comparison with linear and optimal detection schemes for MIMO systems. Most important result is that while it was proven by the analytical results that the ZFML detector is better in large search space, which increases the computational complexity, the SDR detector is computationally efficient detector in same scenario. Possible future work is to analyze SDP and ZFML, BER/computational complexity performance in MIMO system using higher order constellation in a flat fading/Rayleigh channel.

REFERENCES

- [1] Kapila C. Wavegedara, Gaurav Bansal, Wireless Communications: Trends and Challenges chapter from book Next-Generation Wireless Technologies: 4G and Beyond (pp.3-5) Jan 2013.
- [2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020, White Paper, Cisco, Feb. 2016. [Online]. Available: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-520862.pdf>
- [3] D. McQueen, The momentum behind LTE adoption, IEEE Communications Magazine, vol. 47, no. 2, pp. 4445, Feb. 2009.
- [4] M. El-Sayed, A. Mukhopadhyay, C. Urrutia-Valdes, and Z. J. Zhao, Mobile data explosion: monetizing the opportunity through dynamic policies and QoS pipes, Bell Labs Technical Journal, vol. 16, no. 2, pp. 7999, Sep. 2011. [Online]. Available: <http://dx.doi.org/10.1002/bltj.20504>

- [5] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, Mobile data offloading through opportunistic communications and social participation, IEEE Transactions on Mobile Computing, vol. 11, no. 5, pp. 821834, May 2012.
- [6] G. Lawton, Machine-to-machine technology gears up for growth, IEEE Computer Magazine, vol. 37, no. 9, pp. 1215, Sep. 2004.
- [7] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, M2M: from mobile to embedded internet, IEEE Communications Magazine, vol. 49, no. 4, pp. 3643, Apr. 2011.
- [8] Y. Zhang, R. Yu, S. Xie, W. Yao, Y. Xiao, and M. Guizani, HomeM2M networks: architectures, standards, and QoS improvement, IEEE Communications Magazine, vol. 49, no. 4, pp. 4452, Apr. 2011.
- [9] Kuldeep Kumar, A comparison of different detection algorithms in a MIMO system, journal of advanced engineering sciences and technologies, volume no. 7, issue No 2, 301-304, 2011.
- [10] Jun Zhang, Runhua Chen Networked MIMO with clustered linear precoding Wireless Communications, IEEE Transactions, volume: 8, Issue: 4, page(s): 1910-1921, August 2008.
- [11] Sajid Bashir, M. Naeem, An Application of Univariate Marginal Distribution Algorithm in MIMO Communication Systems. International journal of communication system, volume 23, issue 1, pages 109-124, January 2010.
- [12] Adnan Ahmed Khan, Symbol Detection techniques in Spatial Multiplexing systems, PhD Thesis December 2008.
- [13] Xiaodong Li, Howard C, Reduced-Complexity Detection Algorithms for Systems Using Multi-Element Arrays, IEEE telecommunication, volume: 2 page(s): 1072-1076, 2000.
- [14] Wing-Kin Ma, Chao-Cheng Su, The Equivalence of Semidefinite Relaxation MIMO Detectors for Higher-Order QAM, IEEE journal of selected topics in signal processing, vol. 3, no. 6, December 2009.
- [15] P.H. Tan and L.K. Rasmussen, The application of semi-definite programming for detection in CDMA, IEEE Journal on Selected Areas in Communications, 19(8):1442-1449, August 2001.
- [16] Ami Wiesel, Eldar Y, Shamai S, Semidefinite relaxation for detection of 16-QAM signaling in MIMO channels, IEEE Signal Processing Letters, 12(9):653656, 2005.
- [17] Nicholas D Sidiropoulos, A Semidefinite Relaxation Approach to MIMO Detector for High-order QAM constellations, IEEE Signal Processing letters, vol.13, page NO.9, September 2006.
- [18] Andrea Goldsmith, MIMO Wireless Communications, Cambridge University Pr.08-Aug-2005.
- [19] C. E. Shannon, A mathematical theory of communication, The Bell-System Technical Journal, vol. 27, no. 3, pp. 379423, Jul. 1948.
- [20] D. Gesbert, S. V. Hanly, H. Huang, S. Shamai (Shitz), O. Simeone, and W. Yu, Multi-cell MIMO cooperative networks: a new look at interference, IEEE Journal on Selected Areas in Communications, vol. 28, no. 9, pp. 13801408, Dec. 2010.
- [21] S. Shamai (Shitz) and A. D. Wyner, Information-theoretic considerations for symmetric, cellular, multiple-access fading channels Part I, IEEE Transactions on Information Theory, vol. 43, no. 6, pp. 18771894, Nov. 1997.
- [22] Attainable throughput of an interference-limited multiple-input multiple-output (MIMO) cellular system, IEEE Transactions on Communications, vol. 49, no. 8, pp. 13071311, Aug. 2001.
- [23] S. Verdú, Computational complexity of optimum multiuser detection, Algorithmica, vol. 4, no. 1-4, pp. 303312, Jun. 1989. [Online]. Available: <http://dx.doi.org/10.1007/BF01553893>
- [24] D. Micciancio, The hardness of the closest vector problem with preprocessing, IEEE Transactions on Information Theory, vol. 47, no. 3, pp. 12121215, Mar. 2001.
- [25] L. C. Hui and K. B. Letaief, Successive interference cancellation for multiuser asynchronous DS/CDMA detectors in multipath fading links, IEEE Transactions on Communications, vol. 46, no. 3, pp. 384391, Mar. 1998.
- [26] X. Wang and H. V. Poor, Wireless Communication Systems: Advanced Techniques for Signal Reception, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall, 2009.
- [27] R. H. Roy and B. Ottersten, Spatial division multiple access wireless communication systems, U.S. Patent 5 515 378, May 7, 1996.

- [28] M. P. Lotter and P. van Rooyen, Space division multiple access for cellular CDMA, in Proc. IEEE 5th International Symposium on Spread Spectrum Techniques and Applications (ISSSTA98), Sun City, South Africa, Sep. 1998, pp. 959-964.
- [29] G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture, Electronics Letters, vol. 35, no. 1, pp. 1416, Jan. 1999.
- [30] H. Jafarkhani, A quasi-orthogonal space-time block code, IEEE Transactions on Communications, vol. 49, no. 1, pp. 14, Jan. 2001.
- [31] H. Jafarkhani, Space-Time Coding: Theory and Practice. Cambridge University Press, 2005.
- [32] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, Semidefinite relaxation of quadratic optimization problems, IEEE Signal Processing Magazine, vol. 27, no. 3, pp. 2034, May 2010.
- [33] Z.-Q. Luo and W. Yu, An introduction to convex optimization for communications and signal processing, IEEE Journal on Selected Areas in Communications, vol. 24, no. 8, pp. 1426-1438, Aug. 2006.
- [34] L. Vandenberghe and S. Boyd, Semidefinite programming, SIAM Review, vol. 38, no. 1, pp. 4995, Mar. 1996.
- [35] S. Boyd and L. Vandenberghe, Convex Optimization. New York, USA: Cambridge University Press, 2004.
- [36] C. Helmberg, F. Rendl, R. J. Vanderbei, and H. Wolkowicz, An interior-point method for semidefinite programming, SIAM Journal on Optimization, vol. 6, pp. 342-361, 1996.
- [37] Z.-Q. Luo and W. Yu, An introduction to convex optimization for communications and signal processing, IEEE Journal on Selected Areas in Communications, vol. 24, no. 8, pp. 1426-1438, Aug. 2006.
- [38] P. H. Tan and L. K. Rasmussen, The application of semidefinite programming for detection in CDMA, IEEE Journal on Selected Areas in Communications, vol. 19, no. 8, pp. 1442-1449, Aug. 2001.
- [39] W.-K. Ma, C.-C. Su, J. Jalden, T.-H. Chang, and C.-Y. Chi, The equivalence of semidefinite relaxation MIMO detectors for higher-order QAM, IEEE Journal of Selected Topics in Signal Processing, vol. 3, no. 6, pp. 1038-1052, Dec. 2009.
- [40] X. Wang, W.-S. Lu, and A. Antoniou, A near-optimal multiuser detector for CDMA channels using semidefinite programming relaxation, in Proc. IEEE International Symposium on Circuits and Systems (ISCAS01), Sydney, NSW, Australia, May 2001, pp. 298-301.
- [41] A near-optimal multiuser detector for DS-SS systems using semidefinite programming relaxation, IEEE Transactions on Signal Processing, vol. 51, no. 9, pp. 2446-2450, Sep. 2003.
- [42] W.-K. Ma, T.N. Davidson, Quasi-maximum-likelihood multiuser detection using semi-definite relaxation with application to synchronous CDMA, IEEE Transaction Signal Processing, 50(4):912-922, April 2002.
- [43] P.H. Tan and L.K. Rasmussen, The application of semi-definite programming for detection in CDMA, IEEE Journal on Selected Areas in Communications, 19(8):1442-1449, August 2001.
- [44] Shaoshi Yang, Lajos H, Semi-definite Programming Relaxation Based-Virtually Antipodal Detection for Gray Coded 16-QAM MIMO Signaling, IEEE Communication, 978-1-4244-9268-8, November 2011.
- [45] M. Grant, S. Boyd, Disciplined convex programming, Global Optimization, Springer, pages 155-210 2006.
- [46] L. Vandenberghe, S. Boyd, Semi-definite programming, SIAM Review, 38:4995, 1996.
- [47] C. Helmberg, F. Rendl, An interior-point method for semi-definite programming. SIAM Journal on Optimization, 6:342-361, 2005.
- [48] Y.E. Nesterov, Quality of semi-definite relaxation for nonconvex quadratic optimization, Technical Report, University Catholique, Belgium, 1997.
- [49] Mobasher, M. Taherzadeh, R. Sotirov, and A. K. Khandani, An Efficient Quasi-Maximum Likelihood Decoding for Finite Constellations, Department of E&CE, University of Waterloo, Tech. Rep. UW-E&CE 2005-01, 2005, available via the WWW site at <http://www.cst.uwaterloo.ca/amin>
- [50] J. Sturm, Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones., Optimization Methods and Software, vol. 11-12, pp. 625-653, 1999.
- [51] S. J. Benson and Y. Ye, DSDP5 User Guide The Dual-Scaling Algorithm for Semidefinite Programming., Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, Tech. Rep. ANL/MCS-TM-255, 2004, available via the WWW site at <http://www.mcs.anl.gov/benson>.
- [52] J. Sturm, Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones., Optimization Methods and Software, vol. 11-12, pp. 625-653, 1999.
- [53] M. Kojima, K. Fujisawa, K. Nakata, and M. Yamashita, SDPA (Semi-Definite Programming Algorithm) Users Manual Version 6.00., Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2-12-1 Oh-Okayama, Meguro-ku, Tokyo, 152-0033, Japan, Tech. Rep., 2002, available via the WWW site at <http://sdpa.is.titech.ac.jp/SDPA/>.
- [54] J. C. Roh and B. D. Rao, Design and analysis of MIMO spatial multiplexing systems with quantized feedback, IEEE Trans. Signal Processing, vol. 54, no. 8, pp. 2874-2886, Aug. 2006.

Sentiment Analysis, Visualization and Classification of Summarized News Articles: A Novel Approach

Siddhaling Urologin

Department of Computer Science

Birla Institute of Technology and Science, Pilani-Dubai,
Dubai, U.A.E.

Abstract—Due to advancement in technology, enormous amount of data is generated every day. One of the main challenges of large amount of data is user overloaded with huge volume of data. Hence effective methods are highly required to help user to comprehend large amount of data. This research work proposes effective methods to extract and represent the data. The summarization is applicable to obtain a brief overview of the text and sentiment analysis can obtain emotions expressed in the text computationally. The combined text summarization and sentiment analysis is proposed on BBC news articles. A pronoun replacement based text summarization method is developed and VADER sentiment analyzer is used to determine sentiment information. The 3-D visualization schemes have been provided to represent the sentiment information. The sentiment analysis and classification are performed on original BBC news articles as well as on summarized articles using classifiers, such as Logistic Regression, Random Forest and Adaboost. On original news articles highest classification rate of 84.93%, using summarization of ratio 25%, 50% and 75% highest classification rates of 78.73%, 83.06% and 83.23%, respectively are observed.

Keywords—Summarization; sentiment analysis; 3-D visualization; sentiment classification

I. INTRODUCTION

Huge amount of data is being generated every day in the form of social media data, various blogs, web sites, Wikipedia, online newspapers, etc. Due to wide spread usages of social media such as Facebook, Twitter, Yahoo! etc. have enormously increased the amount data that has been produced. The Wikipedia alone contain five million articles and thousands of new articles generated every day. There are various online web sites which are publishing newspapers on daily basis. One of the main challenges of huge data is that user gets over loaded with data and requires effective way to absorb the large volume amount of data. Effective data extraction and representation techniques are needed to help user to comprehend huge data. The text summarization is the technique intended to produce a brief overview of the input text and also reduces the amount of data. Moreover, the sentiment analysis is the computational technique, which deduce the user emotion expressed in the text. The sentiment analysis been effective applied to various fields such as product reviews [1], [2], news articles [3], political debate [4], twitter sentiment data analysis [5], [6], stock market [7], [8], etc.

The goal of the text summarization is to obtain a brief summary of the text [9]. This method of text summarizing can be utilized in different applications namely searching documents related to a particular subject and obtain an overview, gather

headline from newspaper articles, assimilate emails, obtain summary of medical information, to produce brief of scientific articles [10], [11] etc. There are various steps involved in text summarization such as topic identification, interpretation and summary generation [12]. A notable work by Bennostein et al. on topic identification is presented in [13] with a frame work for topic identification and applications. Work on Wikipedia graph centrality method for topic identification is presented in [14]. During text interpretation, the meaning of the text is obtained. For text interpretation researchers have focused on various methods such as ontology based interpretation [15] and text interpretation [16]. The goal of text summarization is to generate an abstract or synopsis on single or multiple documents. J. Alan et al. [11] have presented a text summarization method based on novelty detection at the sentence level. Literature review presented in [17] by Lloret et al. have noted that there are two summarization methods: abstraction and extraction. Semantic representation are constructed from text to produce a brief overview in abstraction method [13], [18]. The extractive summarization methods discussed as in [19]–[21] are intend to choose words, sentences and phrases from the given text to obtain the summary. Forming summary based on frequency of words related to the topic has found suitable application in several area [22], [23] of text analysis. It is observed that in a given document the words that are occurring more frequently indicates the subject on which the text is pivoted. Rafael Ferreira et al. [24] have accessed the sentence scoring technique for text summarization. In their work, it was noted that obtaining the frequency of important words and extract sentence to prepare the summary is one of the effective methods. Pronouns are place the holders for proper nouns, which are often used in the text. In the process of filtering and stopword removing, pronouns are also eliminated affecting the frequency of proper nouns. In this research work the summarizing technique is proposed in which, pronouns are replaced at first with proper nouns and then the frequency of words are computed, thereby enhancing the frequency information related to proper nouns to generate an improved version of the text summary.

Sentiment analysis has found applications in healthcare [25], [26], tourism [27], fraud detection [28], finance [29], politics [30], business [31]. There are additional area of applications that are found in [32]. The sentiment analysis of online news articles is presented in [33]. The prediction of positive and negative sentiment on financial news is carried out in [34]. The opinion mining engine for news article is present in [35], which uses the knowledge from ConceptNet and SenticNet. The sentiment classification described in [36]

uses informatics and theoretic approach. A. Mudinas et al. have presented a notable work [37] on lexicon and concept-level sentiment analysis. T. H. A. Soliman et al. [38] have carried out mining of online customer reviews utilizing support vector machines and a similar work on sentiment analysis has been reported in [39] based on As-LDA model. There is an interesting work reported on sentiment analysis based on machine learning techniques in [40]. Sentence level sentiment analysis has been carried out using cloud machine learning techniques in [41]. Sentiment analysis using different types of lexicon dictionaries are listed in [42], [43].

With motivation to help user to comprehend large volume of data, in this research work, summarization on news articles is performed then carried out sentiment analysis and representation. The extractive text summarization method is developed based on [21] to produce a brief overview of news articles. VADER [42] sentiment analyzer is used on original news articles and summarized news articles to deduce sentiment opinion from the text. By using VADER, various sentiment information has been collected as negative, neutral, positive, compound score and count related to sentiment words. Further sentiment information is represented using several visualizations schemes in three dimension such as column plots, surface plots, scatter plots etc. These 3-D visualization methods give a clear and better scheme to portray the sentiment information. Further the sentiment analysis and classification are carried out on original and summarized news articles using classifiers namely Logistic Regression, Random Forest and Adaboost classifiers. The experiments are carried out on BBC news articles and classification performance is tested on 10-fold cross validation. In Section 2 the method of text summarization with pronoun replacement is described, sentiment visualization and classification is presented in Section 3, an example on summarization and sentiment analysis is given in Section 4. Experiment results are presented in Section 5 followed by Section 6, which covers the conclusion.

II. PRONOUN REPLACEMENT BASED TEXT SUMMARIZATION

The text summarization involves in generating a brief summary of given text. Before generating summary, the preprocessing is carried out on the text. The preprocessing involves noise elimination, lowering text, tokenization, identify stopwords such as *that*, *a*, *the*, etc., and removal of them [44]. During preprocessing, pronouns which are place holders for proper nouns are also eliminated. In this research, a summarization technique is developed in which pronouns are replaced with proper nouns and then extractive summarization is carried. In the extractive methods [24] of text summary generation is to look for keywords or the most important words and their frequency in the text. The approach for identifying the important words is to eliminate stopwords and remaining words are taken as important words. As a part of stopword elimination, pronouns are also eliminated, thereby losing the frequency information. In this research the summarization method of [21] is developed as depicted in Fig. 1. For a given input text, Part of the Speech (POS) tagger of [45], [46] is employed to recognize various parts of a sentence. The pronouns are recognized and replaced with proper nouns. The proper noun that is occurring before pronoun and closer to

pronoun is considered to replace the pronoun. However the original input text is used to produce the final text summary.

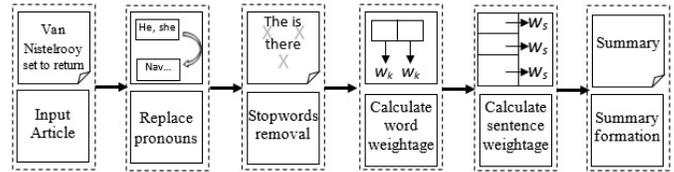


Fig. 1. Pronoun replacement based text summarization.

The next step is to eliminate the stopwords from the text and determine frequency of remaining words in the text. The computation of weightage of keywords or important words is as follows. Let n_k be the number of keywords and n_o be the number of stopwords, then a sentence has total $n_s = n_k + n_o$ words. Let f_k be the frequency of the k^{th} keyword. Also n_t be the number of keywords in the entire text. The weightage of k^{th} keyword can be calculated as

$$w_k = \frac{f_k}{n_t} \quad (1)$$

The sentence weightage is computed as summation of weightage of words given in (2). The sentences having important keywords with more weightage will have higher sentence weightage.

$$w_s = \frac{1}{n_s} \sum_{k=1}^{n_k} w_k \quad (2)$$

The priority order of the sentence is determined using sentence weightage, which indicates order to extract the sentences to form the summary. The user specifies the summary ratio S_r to decide the length of the summary required. For a text with m_t number of lines and S_r given summary ratio, the length of the summary m_s is calculated as

$$m_s = \frac{S_r}{100} \times m_t \quad (3)$$

The text summary is generated by extracting lines in priority ordered up to required length of m_s .

III. SENTIMENT VISUALIZATION AND CLASSIFICATION

The VADER of [42] is a simple rule based sentiment analyzer. It consists of list of lexical features and associated sentiment measures. Based on grammatical and syntactical usage of the language, several rules are formed, which are used to determine the sentiment of the text. A lexicon basically is a list of words with each word assigned a semantic oriented values as positive value or negative value [47]. In VADER list of lexicons the features are assigned values between the range of -4 to +4, here -4 being extreme negative and +4 is extreme positive. In Table I, few words from VADER lexicon list are shown.

It is interesting to perform the sentiment analysis of the news article. Sentiment analysis on news articles are carried in various research such as [33]–[35]. The news article written by an author or journalist provides an opinion on the subject about which article was written. The sentiment analysis thus

TABLE I. EXAMPLE FROM VADER LEXICON

Word	Sentiment Score
Excel	2.0
Exhaust	-1.2
Favorable	2.1
Impatience	-1.8

provides a sentiment evaluation of the news articles. In this research VADER is utilized to perform the sentiment analysis of the BBC news articles. Schematic diagram for sentiment analysis is depicted in Fig. 2. The news articles are subjected to preprocessing such as word tokenization and stopwords removal. Then VADER is applied to compute sentiment score of the news article. The VADER utilizes lexicon list and computes sentiment information such as compound, neutral, negative and positive scores. Also it gives count of positive, negative and neutral words. In this research, novel 3-D visualizations of sentiment information obtained from VADER are presented. The visualizations schemes in terms of three dimensions column plots, surface, scatter plots etc., are developed. These 3D visualization provide better insight of sentiment information gathered from news articles.

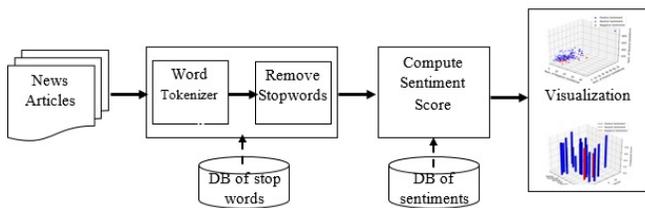


Fig. 2. Sentiment analysis of news articles.

Furthermore, the sentiment classification on summarized news articles as shown in Fig. 3 is performed. As significant amount of data being generated every day, it is becoming important to provide techniques which help user to effectively comprehend the data. The text summarization provide a brief overview of input text and effectively enable user to focus on reduced version of the text. When text summarization is applied to news articles it gives a brief overview of the news with inherent subjective information. Usually the news articles provide elaborated discussion on the subject and hence it is appropriate to perform text summarization to obtain important discussions in news. The sentiment analysis and classification on summarized version of news articles is introduced as shown in Fig. 3. The preprocessing of news articles is performed in which words are tokenized and stopwords are removed. The news articles are subjected to summarization to generate overviews. The sentiment classification is performed on the summarized version of the news articles. Feature vectors of N-gram size are created using a bag of words [48]. Next Logistic Regression, Random Forest and Adaboost classifiers are used for classification. Logistic Regression used as base classifier in Adaboost classifier.

IV. SUMMARIZATION AND SENTIMENT ANALYSIS EXAMPLE

The summarization and sentiment analysis is briefly explained with an example in this section. An input news article is shown in the Fig. 4, which is on Football from BBCSport

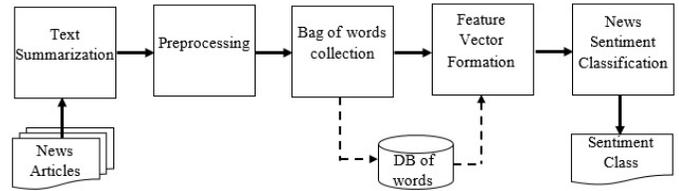


Fig. 3. News summarization and sentiment classification.

news dataset. The number of occurrences of various nouns in this article is determined. Top three nouns with maximum occurrences are ‘Van’, 5 times, ‘Nistelrooy’, 5 times and ‘United’, 4 times. The sentiment score for the Fig. 4 is computed using VADER. Positive score is 0.212, negative score is 0.083, neutral score is 0.705 and compound score is 0.9726 observed.

In this article, the pronoun such as ‘he’, ‘it’, ‘they’ etc., have been used several times as place holder for proper nouns. Text summarization is performed on this text using pronoun replacement method as described in Section 2. Once the pronouns are replaced, the stopwords are eliminated from the text to identify important words or keywords. Next, weightage of each keywords is computing by using (1). Table II shows the computed weightage for the few keyword words from input article.

TABLE II. KEYWORDS AND WEIGHTAGE

Sl. Num.	Keyword	Weightage
1.0	Nistelrooy	0.05625
2.0	United	0.03125
3.0	League	0.0125
4.0	Champions	0.00625
5.0	Fifth	0.00625

The sentence weightage is computed using (2) based on the weightage of keywords present in that sentence. Also each sentence is assigned a priority number based on its sentence weightage. Lower priority number is assigned for the sentence with higher weightage. In Table III sentence (partially depicted), its weightage and priority number for few sentences are shown.

TABLE III. SENTENCE WEIGHTAGE AND PRIORITY NUMBER

Sl. Num.	Sentence	Weightage	Priority Number
1.0	Van Nistelrooy set to return.	0.0225	1.0
2.0	Manchester United striker Ruud van Nistelrooy may make his comeback after an Achilles tendon...	0.0127	5.0
3.0	He has been out of action for nearly three months and had targeted a return in the Champions?	0.0063	12.0
4.0	But Manchester United manager Sir Alex Ferguson hinted he may be back early.	0.0159	4.0
5.0	He said: "There is a chance he could be involved at Everton but we'll just have to see how he comes?"	0.0106	6.0

User provides the summarization ratio, using which the summary is generated. The number sentences to be included in the summary can be found by equation (3) using. The summary is formed by extracting the sentences from the article in priority order. Summarized text for Fig. 4 is collected with

Van Nistelrooy set to return. Manchester United striker Ruud van Nistelrooy may make his comeback after an Achilles tendon injury in the FA Cup fifth round tie at Everton on Saturday. He has been out of action for nearly three months and had targeted a return in the Champions League tie with AC Milan on 23 February. But Manchester United manager Sir Alex Ferguson hinted he may be back early. He said, "There is a chance he could be involved at Everton but we'll just have to see how he comes through training." The 28-year-old has been training in Holland and Ferguson said, "Ruud comes back on Tuesday and we need to assess how far on he is". The training he has been doing in Holland has been perfect and I am very satisfied with it." Even without Van Nistelrooy, United made it 13 wins in 15 league games with a 2-0 derby victory at Manchester City on Sunday. But they will be boosted by the return of the Dutch international, who is the club's top scorer this season with 12 goals. He has not played since aggravating the injury in the 3-0 win against West Brom on 27 November. Ferguson was unhappy with Van Nistelrooy for not revealing he was carrying an injury. United have also been hit by injuries to both Alan Smith and Louis Saha during Van Nistelrooy's absence, meaning Wayne Rooney has sometimes had to play in a lone role up front. The teenager has responded with six goals in nine games, including the first goal against City on Sunday.

Fig. 4. Input news article.

ratio as 25%, 50% and 75% and results of summary are shown in Table IV.

TABLE IV. NEWS SUMMARIZATION AND SENTIMENT ANALYSIS

S_r	Summarized Text	Num. of sentences in summary	Negative Score	Neutral Score	Positive Score	Compound Score
0.25	Van Nistelrooy set to return. But Manchester United manager Sir Alex Ferguson hinted Ferguson may be back..	4.0	0.168	0.749	0.084	-0.4215
0.5	Van Nistelrooy set to return. But Manchester United manager Sir Alex Ferguson hinted Ferguson may be back..	8.0	0.084	0.774	0.142	0.6908
0.75	Van Nistelrooy set to return. But Manchester United manager Sir Alex Ferguson hinted Ferguson may be back..	12.0	0.079	0.726	0.195	0.9618

Further, sentiment analysis using VADER is performed for each summarized text. The VADER computes sentiment information such as negative, neutral, positive and compound score which are given in Table IV.

V. EXPERIMENTAL RESULTS

The experiments are conducted on news article collected from [49], which are BBC articles. The BBCSport dataset includes 737 documents about articles on five topical areas as Athletics, Cricket, Football, Rugby and Tennis from BBC sport web site between the years 2004 to 2005. It is interesting to perform the sentiment analysis on news articles. VADER sentiment analyzer is applied on the news articles on dataset BBCSport. Moreover the POS tagger of [45], [46] is utilized to determine various parts of sentences. Proper nouns and their occurrences in article are gathered. In Table V, top three nouns having maximum occurrence in the articles with their frequency are shown. The VADER sentimental analyzer gives various scores such as negative, neutral, positive and compound score which are given in columns 3, 4, 5 and 6

respectively in Table V. The count of negative, neutral and positive words are given in column 7, 8 and 9, respectively.

Sentiment Visualization:

A novel 3-D visualization of sentimental information obtained from VADER is presented in Fig. 5. Twenty news article on Football and Athletic from BBCSport dataset are considered. For each article, the number of occurrences of proper nouns is determined. In Fig. 5(a) negative sentiment score versus positive sentiment score for each article is represented. In this figure, along x-axis the proper noun with maximum frequency, along y-axis the negative score and along z-axis the positive sentiment score are shown. Fig. 5(b) shows maximum occurring proper noun and count of that noun along x-axis and y-axis against compound sentiment score along z-axis. Fig. 5(b) highlights the compound score on an article with respect to the noun having maximum occurrence and its count, hence showing the importance of the noun as a subject in that article. Fig. 5(c) provides 3D visualization of negative score versus positive score for Athletic articles. In Fig. 5(d) noun occurrences versus compound score is represented for 20 Athletic articles.

The novel 3-D visualizations are developed to represent the compound sentimental score as shown in Fig. 6. In these figures, compound score versus count of positive and negative sentiment words are shown. Fig. 6(a) show the 3-D representation for compound sentiment of all Football articles from the BBCSport dataset. In Fig. 6(a) highest compound score of 2.927 having the number of negative words 6 and number of positive words 14 is observed. Fig. 6(b), 6(c) and 6(d) represent the compound scores for articles on Cricket, Athletic and Rugby respectively are shown. These 3-D visualizations signify the changes in compound score that can occur when count of positive or negative sentiment words vary.

The news articles are subjected to sentiment analyzer VADER, which provides various sentiment score also it gives count of negative, positive and neutral words in the articles. The sentiment of the article is positive for compound score greater than zero, neutral for compound score of zero otherwise it is negative. Fig. 7 provides 3-D visualization of count information obtained from VADER. In Fig. 7(a) the 3-D scatter

TABLE V. NEWS ARTICLE WITH NOUN FREQUENCY IN THE ARTICLE ALONG WITH SENTIMENT SCORES

Article	Noun and frequency	Negative Score	Neutral Score	Positive Score	Compound Score	# Negative Sentiment Words	# Neutral Sentiment Words	# Positive Sentiment Words
1.0	('Everton', 9) (('United', 8) (('Martyn', 8) (('Van', 5)	0.102	0.676	0.221	0.995	19.0	292.0	32.0
2.0	('Nistelrooy', 5) (('United', 4) (('Moyes', 5)	0.083	0.705	0.212	0.9726	6.0	132.0	14.0
3.0	('Beattie', 5) (('Gallas', 5) (('Ronaldo', 3)	0.084	0.805	0.111	0.4504	8.0	204.0	13.0
4.0	('United', 3) (('Manchester', 1) (('Home', 5)	0.044	0.688	0.268	0.9661	2.0	72.0	11.0
5.0	('Smith', 4) (('Scotland', 4)	0.069	0.749	0.182	0.9456	4.0	121.0	11.0

plot is depicted for news articles of Football. More positive sentiments are observed in Fig. 7(a) than negative or neutral. The 3-D scatter plots for Cricket shown in Fig. 7(b), Athletic in Fig. 7(c) and Rugby in Fig. 7(d).

In Fig. 8, ten words with positive sentiment and in Fig. 9, ten words with negative sentiments are depicted. In each graph, the word with its sentiment score and its percentage contribution are shown. In Fig. 8(a) the graph shows positive sentiment words for Football articles. In Fig. 8(b), 8(c) and 8(d) showing words with positive sentiment for Cricket, Athletic and Rugby news articles. Fig. 9 depicts top ten negative sentiment words for news articles.

Summarization and Classification: The sentiment classification is carried out on news articles. The BBCSport news article dataset consists of 737 article related to Football, Cricket, Athletic, Rugby, and Tennis. Later each article is subjected to summarization with ratio of 25%, 50% and 75% hence dataset consists of 2948 articles. The sentiment analysis is performed on each article using VADER. The Logistic Regression, Random Forest and AdaBoost classifiers are used for sentiment classification. Feature vectors of N-gram size are constructed from news articles by preparing bag of words as given in [48]. From the BBCSport dataset of articles occurrences of words are collected and bag of words is prepared by taking 'N' most frequent words. Here 'N' is taken as 2000, 3000 and 4000. Table VI shows 10-fold cross validation results on the BBCSport dataset of articles without summarization. A maximum classification rate of 84.93% is observed for AdaBoost classifier with N as 3000.

Next, the news articles are subjected to summarization using method described in Section 2. The summarization ratio of 25%, 50%, 75% is applied on each article. Using the sentiment analyzer VADER, the sentiment type of each article is determined. The 10-fold cross validations are performed on three classifiers Logistic Regression, Random Forest and AdaBoost classifier with varying 'N' as 2000, 3000 and 4000. In Table VII, the 10-fold cross validation results are presented. It is observed that as summarization ratio increases better sentiment classification rates are obtained. When summarization ratio is 25%, a maximum classification rate of 78.73% for AdaBoost classifier with 'N' equal to 4000 is observed. For summarization ratio of 50%, maximum classification rate 83.06% with 'N' 4000 on AdaBoost classifier is obtained. A maximum classification rate of 83.23% for AdaBoost classifier with 'N' 3000 is observed for 75% text summarization.

VI. CONCLUSION

In the recent years we are witnessing significant amount of data being generated in numerous forms such as social media, web blogs, web sites, Wikipedia, news articles and many more. Due this the end user is overloaded with data and there is a greater need for effective methods to help user to absorb the data. Data extraction and representation methods are highly desirable to assist user to comprehend the huge data. One of the effective methods to obtain brief overview is using text summarization. Also sentiment analysis and classification being used to determine opinion expressed in the text. In this research, the text summarization and sentiment analysis on BBC news articles is combined. BBC news articles are collected from [49], which consists of 737 news articles on various sports topics. Extractive based text summarization method is developed in this research which involves pronoun replacement with proper noun and form text summary. The sentiment analysis of BBCSport news articles is carried by VADER. The VADER provides various evaluated information including positive, compound, negative and neutral score along with count of neutral, negative and positive words in the text. Novel three dimensional visualizations are provided to depict sentiment information obtained on BBCSport. Later, using the summarization ratio of 25%, 50% and 75% the text summarization is carried out on news articles. On the dataset of news articles, the feature vector is formed using bag of words of N-gram size. The sentiment classification is carried out on news articles at first without summarization and later on summarized text of 25%, 50% and 75% ratio. Three classifiers are employed to perform sentiment classification such as Logistic Regression, Random Forest and Adaboost classifier with varying N as 2000, 3000 and 4000. When classification is carried out without summarization highest classification rate of 84.93% observed. For 25%, 50% and 75% summarized text a maximum classification rate of 78.73%, 83.06% and 83.23% are respectively obtained.

ACKNOWLEDGMENT

Author is grateful to Dr. Sunil Thomas, Department of Electrical and Electronics Engg, Birla Institute of Technology and Science, Pilani-Dubai, Dubai for his suggestions to improve the manuscript.

REFERENCES

- [1] Xing Fang and Justin Zhan, Sentiment analysis using product review data, in Journal of Big Data, 2:5, 2015.

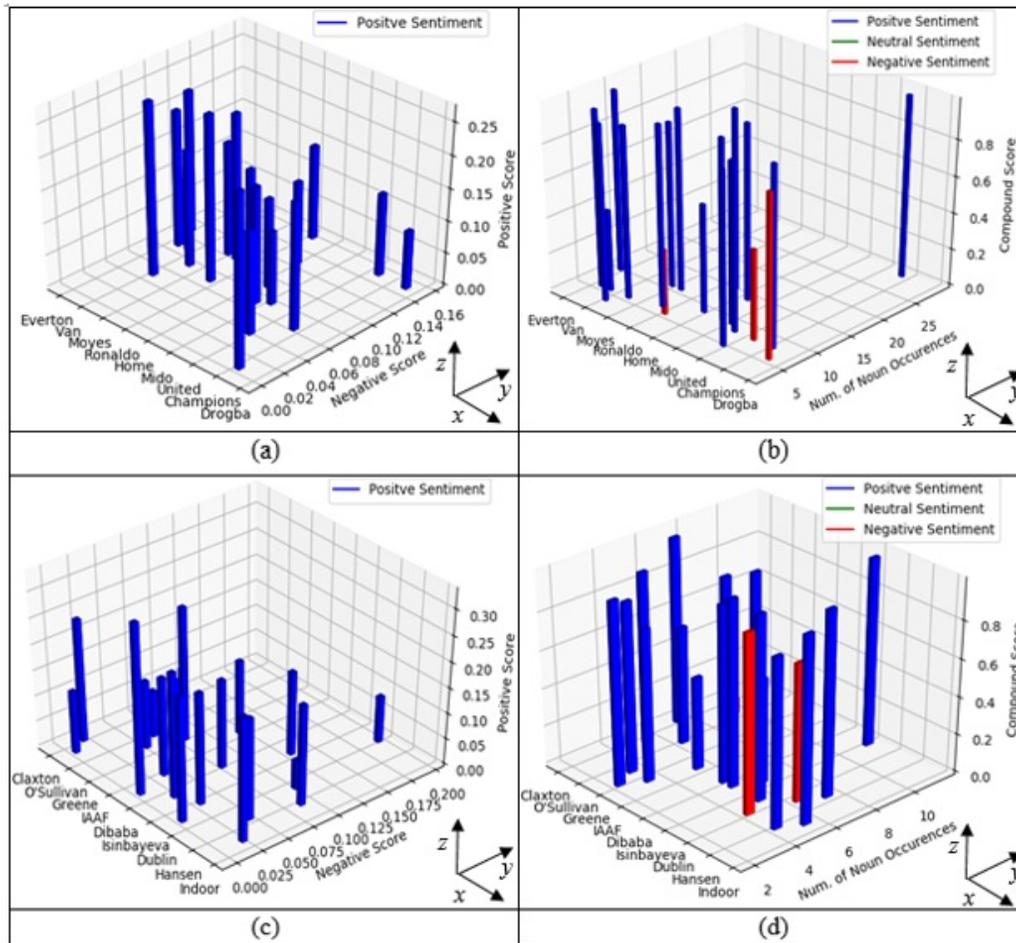


Fig. 5. The 3-D representation of sentimental information for 20 news articles. (a) Negative versus positive score for Football articles. (b) Noun occurrences versus compound score for Football articles. (c) Negative versus positive score for Athletics articles. (d) Noun occurrences versus compound score for Athletics articles.

TABLE VI. PERFORMANCE OF SENTIMENT CLASSIFICATION

	Logistic Regression			Random Forest			AdaBoost		
	2000	3000	4000	2000	3000	4000	2000	3000	4000
N	84.3	84.73	84.3	81.81	81.33	81.02	84.43	84.93	84.63
Classification Rate in %	15.7	15.27	15.7	18.19	18.67	18.98	15.57	15.07	15.37
Misclassification Rate in %									

[2] Himmat M., Salim N., Survey on Product Review Sentiment Classification and Analysis Challenges In: Herawan T., Deris M., Abawajy J. (eds) Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Lecture Notes in Electrical Engineering, vol 285, Springer, Singapore, 2014.

[3] Khoo, C.S.G., Nourbakhsh, A., & Na, J.C., Sentiment analysis of online news text: A case study of appraisal theory Online Information Review, 36(6), 2012.

[4] Yu Wang, Tom Clark, Jeffrey Staton, Eugene Agichtein Towards Tracking Political Sentiment through Microblog Data in Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pages 8893, Baltimore, Maryland USA, 27 June 2014.

[5] Aliza Sarlan, Chayanit Nadam, Shuib Basri, Twitter sentiment analysis, in International Conference on Information Technology and Multimedia (ICIMU), Malaysia November 18 20, 2014.

[6] Youngsub Han, Kwangmi Ko Kim, Sentiment analysis on social media using morphological sentence pattern model, in IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), London, UK, 2017.

[7] Smailovic J., Grear M., Lavrac N., nidaric M., Predictive Sentiment Analysis of Tweets: A Stock Market Application, In: Holzinger A., Pasi G. (eds) Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science, vol 7947. Springer, Berlin, Heidelberg, 2013

[8] Rajat Ahuja, Harshil Rastogi, Arpita Choudhuri, Bindu Garg Stock market forecast using sentiment analysis, in 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pages 1008-1010, 2015.

[9] Fang Chen, Kesong Han and Guilin Chen, An Approach to sentence selection based text summarization, in Proceedings of IEEE TENCON02, pp. 489-493, 2002.

[10] Amari, S.-I. and Nagaoka, H. Methods of Information Geometry, Translations of Mathematical Monographs, in Oxford University Press, 2001.

[11] J. Allan, C. Wade, and A. Bolivar, Retrieval and novelty detection at the sentence level, in Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 314321, 2003.

[12] Hovy, E. and C.-Y. Lin, Automatic Text Summarization in SUMMARIST, in I. Mani and M. Maybury (eds), Advances in Automatic Text

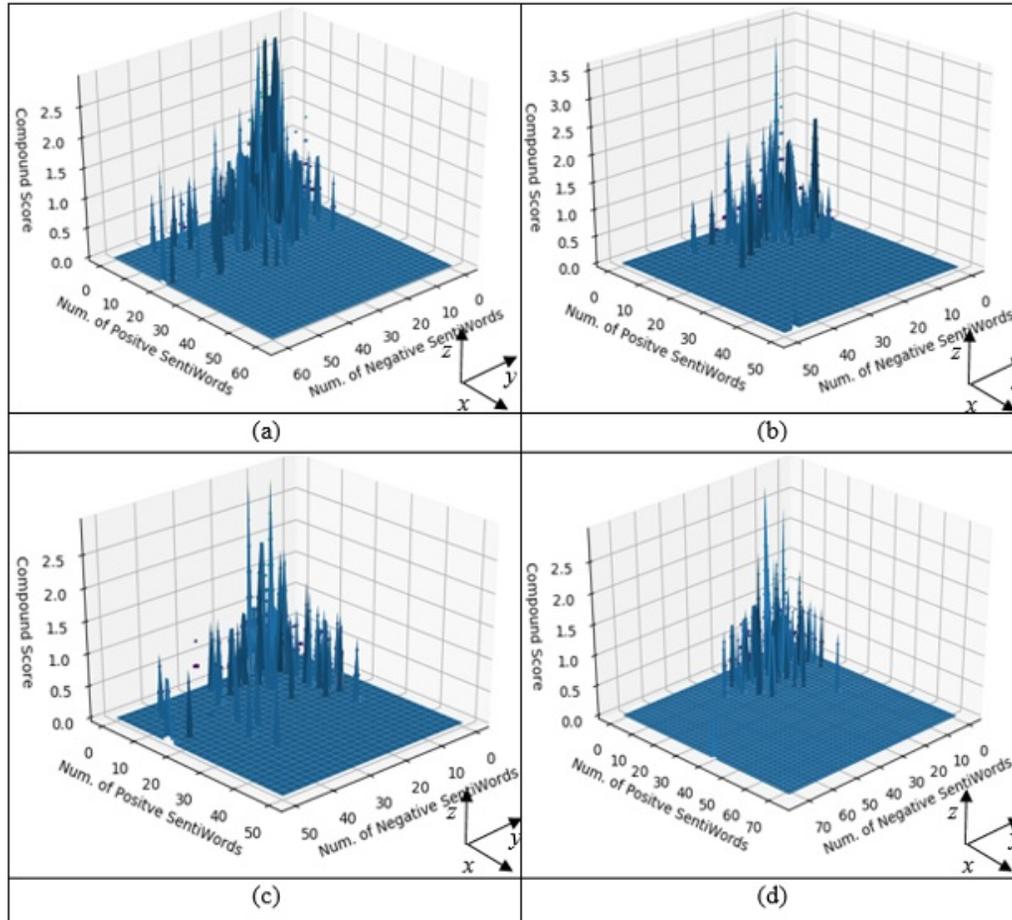


Fig. 6. The 3-D representation of compound score versus count of positive and negative sentiment words. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.

TABLE VII. PERFORMANCE OF SENTIMENT CLASSIFICATION ON SUMMARIZED TEXT

S_r	N	Logistic Regression			Random Forest			AdaBoost		
		2000	3000	4000	2000	3000	4000	2000	3000	4000
0.25	Classification Rate in %	78.5	78.49	78.33	77.24	75.96	76.61	77.57	77.41	78.73
	Misclassification Rate in %	21.5	21.51	21.67	22.76	24.04	23.39	22.43	22.59	21.27
0.5	Classification Rate in %	82.73	81.06	81.31	79.76	78.98	80.22	82.1	82.06	83.06
	Misclassification Rate in %	17.27	18.94	18.69	20.24	21.02	19.78	17.9	17.94	16.94
0.75	Classification Rate in %	82.16	82.43	82.12	81.65	79.74	79.77	82.71	83.23	82.2
	Misclassification Rate in %	17.84	17.57	17.88	18.35	20.26	20.23	17.29	16.77	17.8

Summarization, pp.81-94, MIT Press, 1999.

- [13] Benno Stein, Sven Meyer zu Eissen, Topic Identification: Framework and Application, in Proceedings of International Conference on Knowledge Management, pp 522-531, 2004.
- [14] Kino Coursey, Rada Mihalcea, Topic Identification Using Wikipedia Graph Centrality, in Proceedings of NAACL HLT 2009, pages 1171-120, Boulder, Colorado, June 2009.
- [15] Irma Sofia Espinosa Peraldi, Atila Kaya, Sylvia Melzer, Ralf Moller, On Ontology Based Abduction For Text Interpretation, in Proceedings of 9th International Conference Computational Linguistics and Intelligent Text Processing, Israel, pp 194-205, 2008.
- [16] Marti A. Hearst, Direction-Based Text Interpretation as an Information Access Refinement, in Text-Based Intelligent Systems, Lawrence Erlbaum Associates, 1992.
- [17] Lloret, E. & Palomar, M, Text summarisation in progress: a literature review, in Artificial Intelligence Review, vol. 37, issue 1, pp 1-41, January 2012.
- [18] Kazuo Sumita, Seiji Miike, Kenji Ono, Tetsuro Chino, Automatic abstract generation based on document structure analysis and its evaluation as a document retrieval presentation function, in Systems and Computers in Japan, vol. 26, issue 13, 2007.
- [19] Zhang Pei-ying and LI Cun-he Automatic text summarization based on sentences clustering and extraction, in 2nd IEEE International Conference on Computer Science and Information Technology, pp. 167-168, 2009.
- [20] Daniel Gayo-avello , Daro Ivarez-gutierrez , Jos Gayo-avello, Naive Algorithms for Key-phrase Extraction and Text Summarization from a Single Document inspired by the Protein Biosynthesis Process, in First International Workshop Biologically Inspired Approaches to Advanced Information Technology, BioADIT 2004, Lausanne, Switzerland, pp. 440-455, January 29-30, 2004.
- [21] Siddhaling Urolagin, Jagadish Nayak, Likitha Satish A method to generate text summary by accounting pronoun frequency for keywords weightage computation, in Proceedings of 2017 International Conference on Engineering & Technology (ICET'2017) Akdeniz University, Antalya, Turkey, pages 1-4, 21-23 August, 2017.
- [22] Jos M. Perea-Ortega, Elena Lloret, L. Alfonso Urea-Lpez, Manuel Palo-

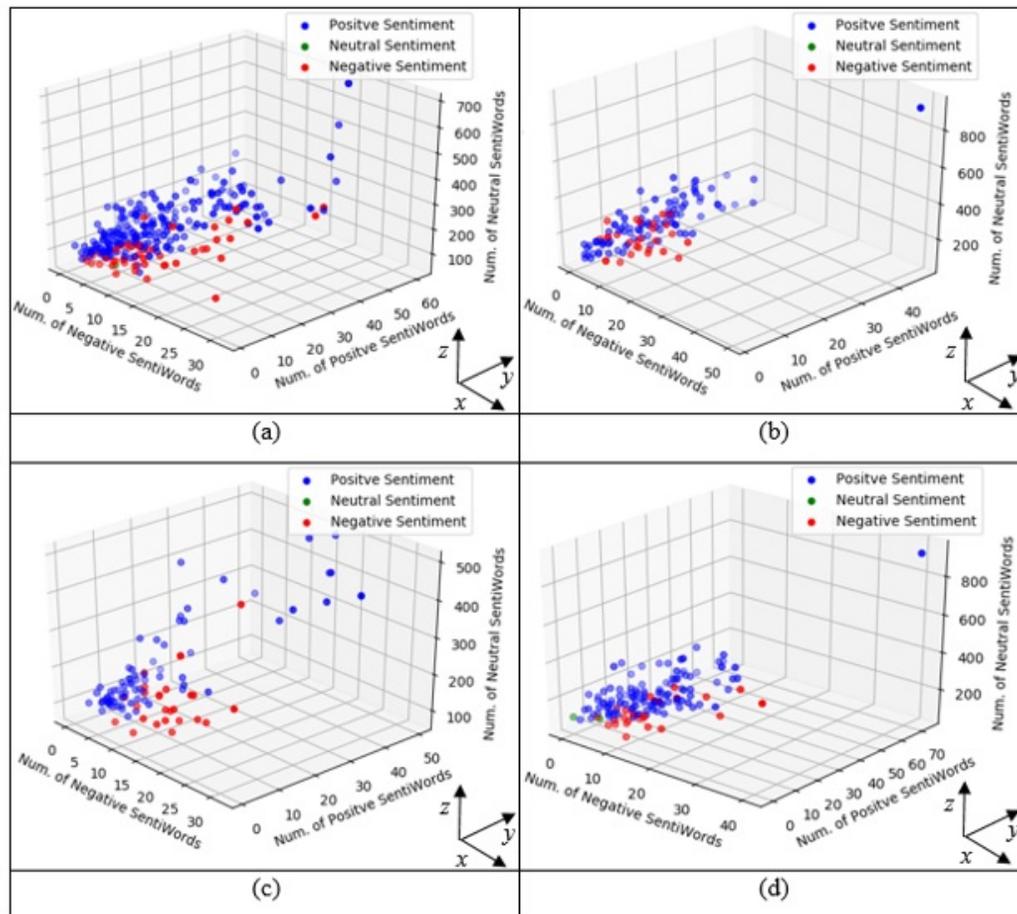


Fig. 7. Representing sentimental count information as 3-D visualization. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.

mar, Application of Text Summarization techniques to the Geographical Information Retrieval task, in Expert Systems with Applications, vol. 40, issue 8, pp. 29662974, 2013.

[23] Nongnuch Ketui, Thanaruk Theeramunkong, Chutamane Onsuwan, An EDU-Based Approach for Thai Multi-Document Summarization and Its Application, in Journal ACM Transactions on Asian and Low-Resource Language Information Processing TALLIP, vol. 14 issue 1, January 2015.

[24] Rafael Ferreira, Luciano de Souza Cabral, Rafael Duerre Lins, Gabriel Pereira e Silva, Fred Freitas, George D.C. Cavalcanti, Rinaldo Lima, Steven J. Simske, Luciano Favaro, Assessing sentence scoring techniques for extractive text summarization, in Expert Systems with Applications, vol.40, issue 14, pp. 57555764, 2013.

[25] Vincenza Carchiolo, Alessandro Longheu, and Michele Malgeri, Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics, in Proceedings of the 6th International Conference on Information Technology in Bio- and Medical Informatics - Volume 9267, pp. 16-24, 2015.

[26] Denecke K, Sentiment Analysis from Medical Texts. In: Health Web Science. Health Information Science, in Springer, Cham. doi:https://doi.org/10.1007/978-3-319-20582-3-10, 2015.

[27] D Grabner, M Zanker, G Fliedl, M Fuchs, Classification of Customer Reviews based on Sentiment Analysis, in Information and Communication Technologies in Tourism pp 460-470. 2012.

[28] Gann W-JK, Day J, Zhou S, Twitter analytics for insider trading fraud detection system in Proceedings of the second ASE international conference on Big Data. ASE. May 27 - May 31, Stanford, CA, USA, 94305, 2014.

[29] Siddhaling Urolagin, Text Mining of Tweet for Sentiment Classification and Association with Stock Prices, in 2017 International Conference on Computer and Applications (ICCA), Doha, 2017, pp. 384-388. doi: 10.1109/COMAPP.2017.8079788.

[30] Kartik Singhal, Basant Agrawal, Namita Mittal, Modeling Indian General Elections: Sentiment Analysis of Political Twitter Data, Advances in Intelligent Systems and Computing, vol 339. Springer, New Delhi, pp 469-477, 2015.

[31] Van Looy A., Sentiment Analysis and Opinion Mining (Business Intelligence 1), In: Social Media Management. Springer Texts in Business and Economics. Springer, Cham, 2016.

[32] M.Walaa, A. Hassan, and H. Korashy, Sentiment Analysis Algorithms and Applications: A Survey, Ain Shams Engineering Journal, vol.5, no. 4, pp. 10931113, 2014.

[33] Simon Fong, Yan Zhuang, Jinyan Li, Richard Khoury, Sentiment Analysis of Online News Using MALLETT, International Symposium on Computational and Business Intelligence (ISCBI), 2013.

[34] Bradley Meyer, Marwan Bikdash, Xiangfeng Dai, Fine-Grained Financial News Sentiment Analysis, in SoutheastCon, pages 1-8, At Charlotte, NC, USA, 2017.

[35] Prashant Raina Sentiment Analysis in News Articles Using Sentic Computing, in IEEE 13th International Conference on Data Mining Workshops (ICDMW), 2013.

[36] Lin Y, Zhang J, Wang X, Zhou A, An information theoretic approach to sentiment polarity classification, in Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality, ACM, New York, NY, USA, pp 3540, 2012.

[37] A. Mudinas, D. Zhang and M. Levene, Combining lexicon and learning based approaches for concept-level sentiment analysis, in Proceedings of the First International Workshop on Issues of Sentiment Discovery

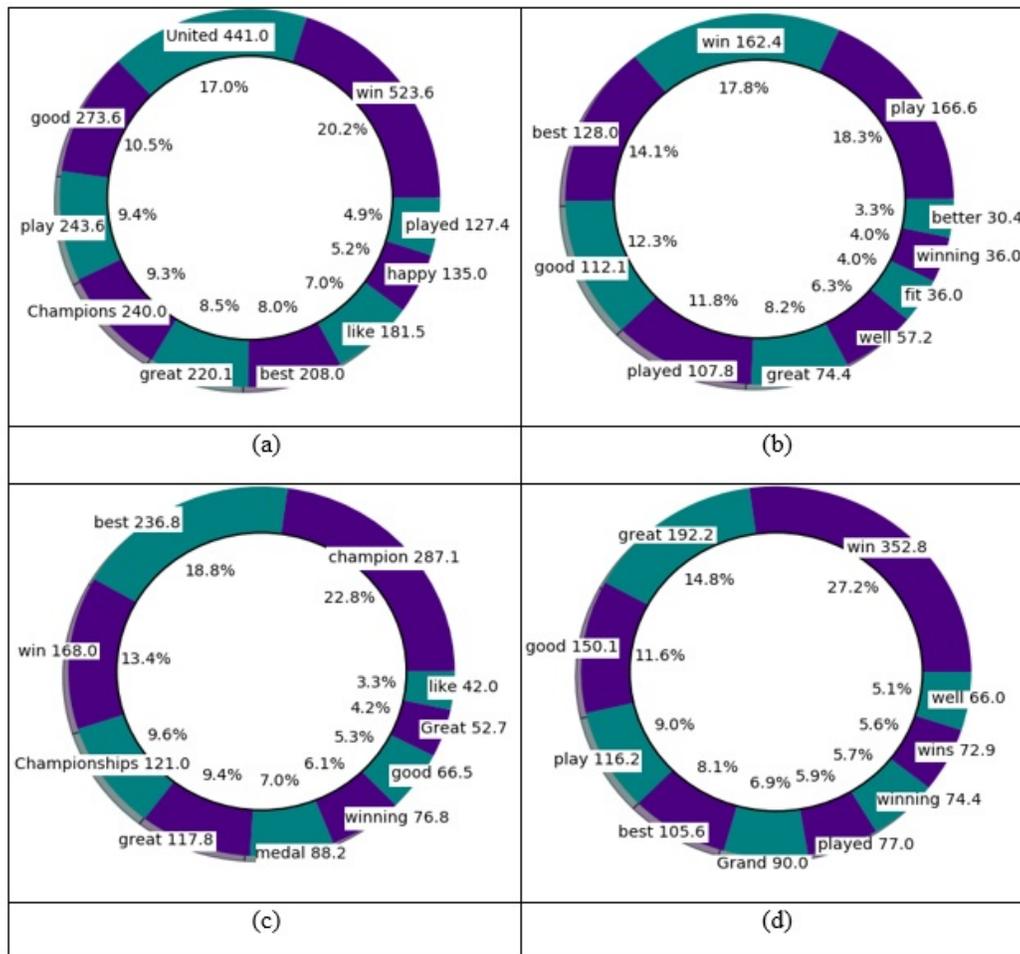


Fig. 8. Top ten positive sentiment words. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.

and Opinion Mining, Beijing, China ugust 12 , 2012, Article No. 5, doi:10.1145/2346676.2346681, 2012.

[38] T. H. A. Soliman, M. A. Elmasry, A. R. Hedar and M. M. Doss, Utilizing support vector machines in mining online customer reviews, in 22nd International Conference on Computer Theory and Applications (ICCTA), Alexandria, 2012, pp. 192-197. doi: 10.1109/ICCTA.2012.6523568, 2012.

[39] J. Liang, P. Liu, J. Tan, and S. Bai, Sentiment Classification Based on AS-LDA Model, *Procedia Computer Science*, vol. 31, pp. 511516, 2014.

[40] A. P. Jain and P. Dandannavar, Application of Machine Learning Techniques to Sentiment Analysis, in 2nd International Conference on Applied and Theoretical Computing and Communication Technology, Bangalore, pp. 628632, 2016.

[41] R. Arulmurugan, K. R. Sabarmathi, and H. Anandakumar, Classification of sentence level sentiment analysis using cloud machine learning techniques, in pages 1-11, *Cluster Computing*, 2017, <https://doi.org/10.1007/s10586-017-1200-1>.

[42] Hutto, C.J. and Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, in Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[43] Y. Wang, Y. Zhang, and B. Liu, Sentiment Lexicon Expansion Based on Neural PU Learning, Double Dictionary Lookup, and Polarity Association, *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, pp. 711, 2017.

[44] Alper Kursat Uysal and Serkan Gunal, The impact of preprocessing on text classification, in *Information Processing & Management* 50, 1, pages 104112, 2014.

[45] Kristina Toutanova and Christopher D. Manning, Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63-70. 2000.

[46] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer, Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, in *Proceedings of HLT-NAACL*, pp. 252-259, 2003.

[47] Liu, B., *Sentiment Analysis and Subjectivity*, in N. In-durkhya & F. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed.). Boca Raton, FL: Chapman & Hall, 2010.

[48] A. Deshwal and S.K. Sharma, Twitter sentiment analysis using various classification algorithms, 5th International Conference on Reliability, Infocom Technologies and Optimization, Noida, pp. 251-257, 2016.

[49] D. Greene and P. Cunningham. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering, in *Proceeding ICML '06 Proceedings of the 23rd international conference on Machine learning* Pages 377-384, 2006.

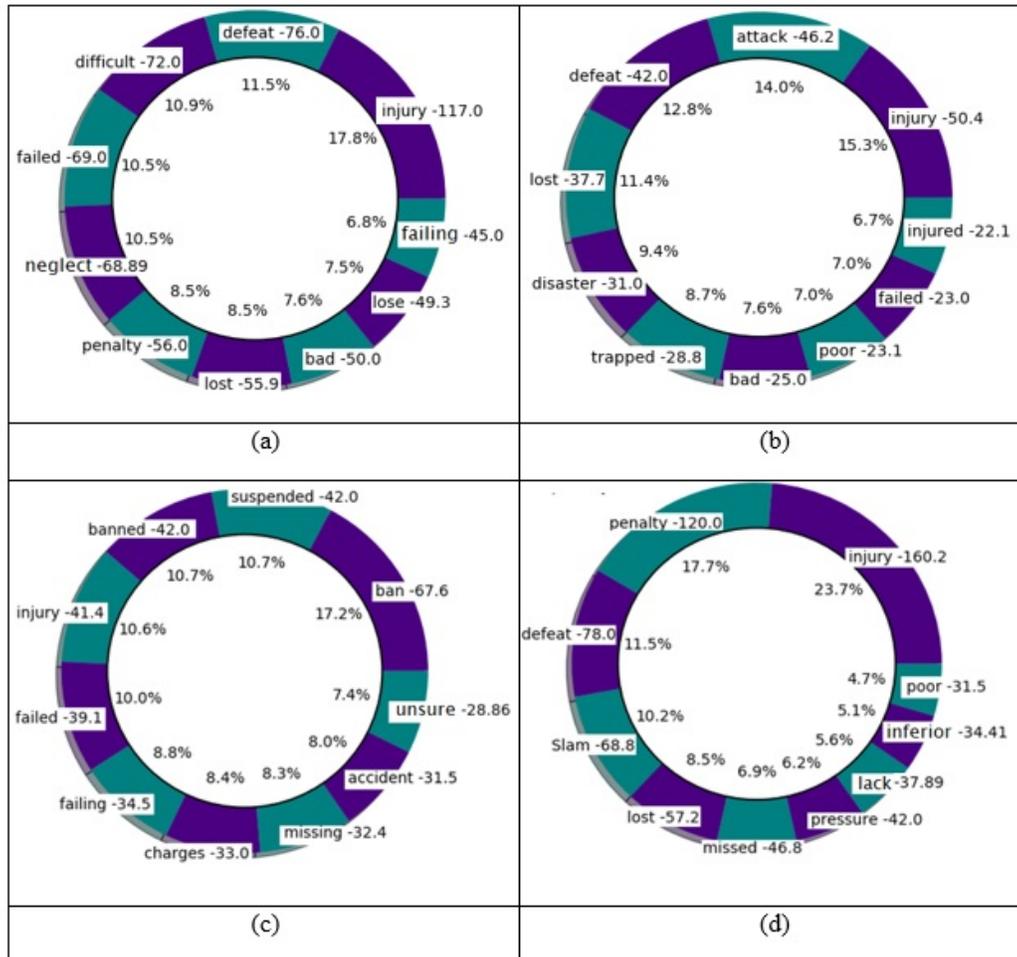


Fig. 9. Top ten negative sentiment words. (a) For article on Football (b) For article on Cricket (c) For article on Athletic (d) For article on Rugby.

Minimization of Information Asymmetry Interference using Partially Overlapping Channel Allocation in Wireless Mesh Networks

Sadiq Shah¹
Department of
Computer Science,
FATA University
FR, Kohat, Pakistan

Khalid Saeed²
Department of Computer Science
Shaheed Benazir Bhutto
University, Sheringal, Dir-Upper Pakistan

Mustafa Khan³, Rafi Ullah Khan⁴,
Mohib Ullah Khan⁵, Misbah Daud⁶,
Arbab Waseem Abbas⁷
The University of
Agriculture, Peshawar, Pakistan

Abstract—Wireless Mesh Network (WMN) is a developing technology that has a great impact on the improvement of the performance, flexibility and reliability over the traditional wireless networks. Using multi-hop communication facility these networks are installed as a solution to extend last-mile access to the Internet. WMN has already been deployed but still it faces certain issues regarding channel assignment and interference. One of the well-known interference issues is Information Asymmetry (IA) interface that results in increased retransmission ratio, end-to-end delay, and thus decreases the overall network capacity of WMN. Various studies have been done in the past to minimize information asymmetry interference using limited number of orthogonal or non-overlapping channels i.e. 1, 6 and 11 from IEEE 802.11b radio technology. In recent studies, it is mentioned that partially overlapping channels called POCs can be used to maximize network capacity. The purpose of this research is to minimize Information Asymmetry (IA) interference problem by proposing a channel assignment model called Optimal Partially Overlapping Channel Assignment (OPOCA). In this research, comparison has been made between OPOCA and existing Information Asymmetry Minimization (IAM) model. Through extensive simulations it has been verified that the proposed OPOCA model gives 8% better results as compared to existing IAM model.

Keywords—Wireless Mesh Network (WMN); information asymmetry; Optimal Partially Overlapping Channel Assignment (OPOCA); NOC; Information Asymmetry Minimization (IAM) model

I. INTRODUCTION

A. Wireless Mesh Network

Wireless Mesh Network (WMN) is a well-known technology that has the capability of better performance regarding flexibility and reliability from that of conventional wireless networks. WMNs have self-healing capabilities and are easier to install. Fixed wireless broadband networks of next generation are being installed increasingly as mesh networks, for the purpose to extend access and provide Internet globally. The wireless networks could be set up between hundreds of wireless mesh nodes that would talk to each other for the purpose to share the data through network across vast area. Nodes of wireless mesh network are small radio transmitters that acts like router (wireless) using current standards of Wi-Fi (802.11 a, b and g) for the purpose of communications. A

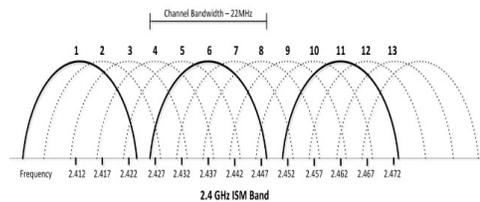


Fig. 1. Non-overlapping and Partially Overlapping Channels (POC) [3].

Wireless Mesh Network contains gateways, mesh clients and mesh routers. Mesh routers forward traffic to the gateways and from the gateways that might connect to the Internet, while the wireless devices, cell phones and laptops are referred as mesh clients [1]. Traditionally the wireless networks were supplied with only one radio interface. Though, the only interface intrinsically limits the overall network using only a single channel. In case of two neighboring links that are operating on the similar channel and transmit data concurrently, then the chances of interference is definitely more with each other. Multi-radio multi-channel (MRMC) systems, the nodes can receive and transmit simultaneous transmission to multiple neighbors can be easily done.

B. Channel Assignment in WMNs

The Channel Assignment (CA) in a multi-radio multi-channel WMN comprises a mission to allocate channels to different radio interfaces in such a technique that achievement of effective channel utilization and minimization of the interference could be made possible. The Wi-Fi standard 802.11b/g works in 2.4 GHz frequency spectrum and has supported transmission capacity of 11Mbps. In IEEE 802.11b spectrum 11 POC channels are used [2]. Those wireless channels that have spectrum overlap with the other working channels are referred as Partially Overlapped Channels (POCs). The Non-Overlapped Channels (NOCs) are those channels that have no spectrum overlap with any other channels that are working. In IEEE 802.11b/g wireless standard which is quite well known, the largest channel that is orthogonal (non-overlapping) set contains channel 1, 6 and 11. One of foremost problems in proposing effective schemes of channel assignment using POCs is the adjacent channel interference that is the interference among two neighbors adjacent. In Fig. 1 IEEE 2.4GHz

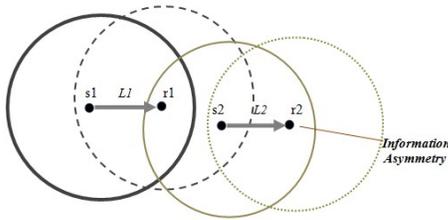


Fig. 2. Information Asymmetry (IA) interference in WMN.

spectrum is shown.

C. Information Asymmetry Interference in WMN

Interference plays a vital role when the number of different clients shares frequency using multiple channels. The interference in WMN is categorized as the coordinated and Information Asymmetry (IA) interference. Fig. 2 shows IA interference scenario. In case of IA the source nodes s_1 and s_2 are outer from the carrier sensing range (CS) of each other and the destination nodes r_1 and r_2 are also outer from the range of each others carrier sensing range CS. In the same way r_2 and s_1 are also located outside CS range of each other but r_1 and s_2 are inside the CS ranges of each other. In interference of Information Asymmetry (IA), the sender node is exterior to the Carrier Sensing range (CS) i.e., if there are two links L_1 (r_1, s_1) and L_3 (r_2, s_2) that are working on similar channel then following condition needs to be accomplished for IA interference [3].

- $d(s_1, s_2) > CS$
- $d(r_1, s_2) < CS$
- $d(s_1, r_2) > CS$

Here d means the distance between physical nodes and CS shows the Carrier Sensing range. In Fig. 2, the carrier sensing range of the sender nodes is represented by solid line circle and the circle with dotted line shows CS range of the receiving nodes.

D. Paper Contributions

To identify IA interference links in MR-MC wireless mesh network. To propose the most optimal channel assignment strategy in MR-MC WMN that will minimize IA interference effect and maximize network capacity in MRMC wireless mesh network. To compare the proposed model results with existing model called Information Asymmetry Minimization (IAM) model using IEEE 802.11b technology. To verify proposed model results using extensive simulations.

II. RELATED WORK

Authors in [4] examined the decline in capacity of multiple radios WMN because of interference the links of communication through wireless network. This method was fundamentally used to reduce the interference among different inter communication links and maximize the network capacity. CA approach which is based on cluster is installed in order to lessen the complication of CA and use again the channel in the complete different cluster. Similarly [5] proposed that the efficient use

of available channels improves the overall performance of a system. Most of the channels are partial overlapped. It is not always unsafe to use the partially overlapped channels. Different radio spectrum is discussed. The author gives overview of different interference constraints that are hard constraint, soft constraint, traffic demand constraint and orthogonal constraint. This paper gives the solution to prevent interference occurred due to partially overlapping.

A detailed survey was presented in [6] regarding survey of some of the channel assignment approaches. In this research the network model and key design concerns are identified. The comparisons of four graph based algorithms have been made that are Breadth First Search Channel Assignment (BFS-CA), Minimum Interference Survivable Topology Control (INSTC), Connected Low Interference Channel Assignment (CLICA), Centralized Tabu-based Algorithm (CTA). According to [7] the solid information about issues concerned to the most favorable usage of radio channels and interfaces in WMNs is examined. The aim of this paper is to provide new central model of channel assignment that is called First Random Channel Assignment algorithm that is associated with the two other CA techniques called Load Aware Channel Assignment (LACA) and Clear Channel Assignment (CCA) by the equal Quality of Service parameters. NS-2 network simulator is used for simulation. In [4] the author compared the results of channel assignment for both dense and sparse MRMC WMN topologies through extensive simulation. The model optimization gives optimal outcomes for the environments where the non-coordinated interference is more than in sparse environments. An algebraic model is formulated that gives strategy of an optimized channel assignment. In this paper the simulation results showed that the intended optimization model performance is 19% better in sparse MRMC WMN topologies where the nCO interference is high [4].

Venkata et al., in [8] recommended two new channel assignment algorithms in this research, BFS-MinNI, BFS-MaxNI to minimize the interference for WMN. The results were compared of both the proposed CA algorithm and the existed algorithms like BFS-CA and CCA. The co-located interference between the radio interfaces and measure the degree of interference value of a variety of grid sizes and evaluated the performance of different algorithms. The experiment based on BFS-MinNICA results in significant performance improvement at varying interference levels. Wang et al., in [9] proposed that the network capacity can be improved by utilizing POCs, which assign channel to all the links while minimizing total network interference. The direct relationship between interference range and channel separation was calculated. The author evaluated that as the network traffic between the internet and clients become prevailing; distance from the gateway, the number of neighboring nodes and interference are used to resolve the channel assignment order of the links. The simulation results showed that the end to end delay and network throughput can be improved by making use of both non orthogonal channels as well as orthogonal channels.

Sadiq et al., [11] proposed an Information Asymmetry Minimization model called IAMin to minimize information asymmetry interference in multi-radio-multi-channel wireless mesh networks. The proposed channel assignment model optimally allocate IEEE 802.11b/g non-overlapping channels to wireless

TABLE I. NOTATION USED IN MODEL

SYMBOL	DESCRIPTION
L	Set of all mesh links
H	Number of channels that is 11
λ	Traffic flow fraction on any link
e_i	A WMN link
c_j	Frequency channel
k	The set of non-overlapping channels that is 3
C_{c_j}	frequency channel capacity
$x(e_i, c_j)$	Channel c_j assigned to link e_i
$IA(e_i)$	Set of IA interfering links of link e_i
$CO(e_i)$	Coordinated interference link of e_i

links of multi-radio multi-channel wireless mesh network. The optimal channel assignment not only minimizes information asymmetry problem, but also maximizes the overall network capacity. Simulation result show that the proposed optimization model successfully minimizes information asymmetry interference and maximizes the capacity in sparse scenarios of multi-radio multi-channel wireless mesh networks up to 8

Another research article [12], compares near-hidden and information asymmetry problems in wireless mesh networks. The comparison between existing Optimal Channel Assignment Model (OCAM) and Information Asymmetry and Near Hidden Minimization (INM) model is done to find the networks capacity. The proposed model considers three non-overlapping channels 1, 6 and 11 from IEEE802.11b standard. An extensive simulation in OPNET shows that the proposed INM model performs 7% better than the existing OCAM model.

III. RESEARCH METHODOLOGY

In this section the optimization model based on partially overlapping channel assignment is proposed to minimize Information Asymmetry (IA) interference problem.

A. Proposing Linear Programming POC model

A linear programming model is proposed that is base on Partially Overlapping Channel (POC) assignment that consists of objective function and several channel assignment constraints. The wireless mesh network allows simultaneous use of multiple channels to increase the aggregate capacity. Partially Overlapped Channels has the potential of increasing the capacity in WMNs by allowing more links to transmit at the same time. Channel assignment using POC allows significantly more flexibility in sharing the wireless spectrum. Here POC model is implemented for maximizing the network throughput and performance using maximum channel resources in a multi-channel multi-radio wireless mesh network.

B. Problem Formulation

In this section we formulate a linear optimization model. The model comprises of decision variable, objective function and certain set of constraints. Table I shows all the symbols and notations used in this model.

C. Existing Model

The IEEE 802.11b standard is used in existing IAMin model. Each networks node is well-equipped with two or

more radio interfaces for taking advantage of multi-radio multi-channels WMN. The existing model is Information Asymmetry Minimization (IAMin) model. This model consist a set of definite constraints, objective function and decision variable.

1) *Decision Variable* : In the decision variable model an IEEE 802.11b non overlapping channel that is c_j is allocated to a link named as e_i . It is stated that if any link that is directed e_i operates on any channel c_j in a network then it equates 1 and if no channel is assigned to any specified link then it equates 0. Such type of a variable (usually decision variable) is also known as binary variable [4].

$$x(e_i, c_j) = \begin{cases} 1 & \text{Directed link } e_i \text{ active on channel } c_j \\ 0 & \text{otherwise} \end{cases}$$

2) *Set of Constraints*: Constraints that are also the restrictions and limitations on an optimization model and they explain the unfavorable results. The interference effects between the links can be minimized for the improvement of the performance of overall network. Following are some set of constraints used for this channel assignment model.

- Individual channel per link constraint

The first constraint of the optimization model presented makes it sure that each and every link in a set of L (edges) should be allocated only one channel from the set H. H is the set of total number of channels. Equation of single channel per link constraint affirms that if is a link that is assigned to a channel than calculation of overall channels evaluates to 1 [4].

$$\sum_{c_j \in H} x(e_i, c_j) = 1 \quad \forall e_i \in L, c_j \in H$$

- Channel Capacity Constraint

If the links that are coordinated are working on similar frequency channel then there is minimum interference and maximum network performance. The coordinated links are not affected by the interference. The capacity of a frequency channel is in fact dispensed amongst all the coordinated interfering links. The rate of traffic at a link should not go beyond the capacity of this link. The constraint that is given below indicates that the capacity of a channel is contributed when numerous corresponding links are active on a similar channel [10].

$$x(e_i, c_j) \cdot \lambda(e_i) \cdot f(e_i) + \sum_{e_k \in CO(e_i)} x(e_k, c_j) \cdot \lambda(e_k) \cdot f(e_k) \leq C_{c_j} \quad \forall e_i \in L, c_j \in H$$

- Information Asymmetry interference Constraints

The IA interference constraint shows that if there are multiple IA interfering links in a network then only one channel c_j will be assigned to them. The equation shows that e_i and e_k are numerous links that are operating on the same channel i.e., c_j in a network. Individual link will be active on channel c_j in the interference region [10]. Fig. 3 shows one of the examples of IA interference.

$$x(e_i, c_j) + \sum_{e_k \in IA(e_i)} x(e_k, c_j) \leq 1$$

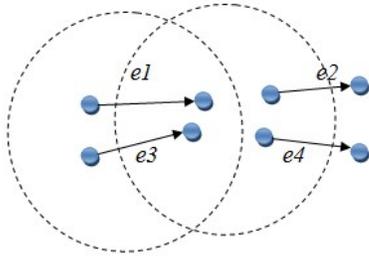


Fig. 3. Information Asymmetry links Fig. 3 illustrates information asymmetry links. The channel c_j is assigned to different links that are e_1, e_2, e_3 and e_4 . The sender nodes are outside each other CS.

$$\forall e_i \in L, \forall c_j \in H, e_k \in L$$

D. Proposed Model

The proposed model consist of node-radio constraint and Partially Overlapping Channel constraints. The proposed model is Optimization Partially Overlapping Channel Assignment (OPOCA) Model.

1) *Node-Radio Constraint*: Such type of constraint relates to nodes of wireless mesh network that comprises of multiple radios. A node in a network can use at most $n(v_i)$ radios in a given period of time for transmission or receiving of data or used for both. This leads towards the subsequent constraint. It is made sure by the constraint that the total quantity of channels that operates on various links of an incident node must not increase than the number of radio interfaces on that specified node [4].

$$\sum_{c_j \in H} \sum_{e_i \in I(v_i)} x(e_i, c_j) \leq n(v_i)$$

$$\forall e_i \in L, \forall c_j \in H, \forall v_i \in V, c_j \in H, e_i \in L$$

2) *Partially Overlapping Channel (POC) Constraint*: Partial overlap among different channels decreases the received flow transmission on one single channel if same channel is assigned to neighboring links. Two nodes can operate on the same channel as long as they do not interfere with each other. The minimization in the interference range increases with increase in separation. The channels, with complete reduction perceived over non-overlapping channels. In Fig. 4 POCs can be in use to interconnect to the number of networks or can be used to add flexibility to the routing infrastructure by creating additional edges in the mesh network topology. In Table II, the channel separation illustrates the difference in channel numbers which is denoted by $cs(c_i - c_j)$. In Fig. 4, there are two communication nodes pairs that are transmitting on a channel c_j and channel c_i . The distance is increased gradually between these two nodes and interference range is recorded.

$$x(e_i, c_j) + \sum_{ek \in IAei[cs|c_i - c_j]} x(ek, c_i) \leq 1$$

$$\forall c_i \in poc(c_j)$$

TABLE II. CARRIER SENSING RANGES [3]

Channel Separation(M)	0	1	2	3	4	5
i=1	13.26	9.08	7.59	4.69	3.21	0
i=6	12.89	9.21	6.98	5.15	3.84	0

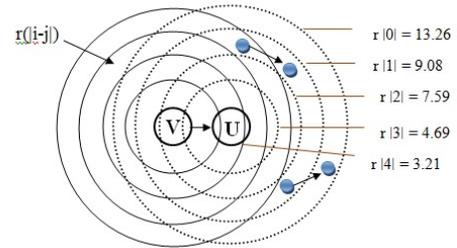


Fig. 4. Node U transmitting on j needs to be in $cs - c_i - c_j$ —to interfere with node V transmitting on channel i [3].

IV. RESULTS AND DISCUSSION

In this section the results that are gathered from simulations are discussed.

A. MATLAB WMN Topology

In MATLAB three different topologies are created for experimental purposes. Each topology consists of 25 mesh nodes. The nodes are aligned close to each other, as the IA interference effect is verified on several numbers of nodes. The greatest Transmission range (Tr) of every node is 10 meters while the carrier sensing range (CS) is assumed to be 20 meters. The Tr is a range where successful communication is occurred among different nodes in the network. All nodes are vigorous and share information among each other.

Fig. 5 and 6 represents MRMC WMN topologies which are created in MATLAB software. In this figure the x-coordinates and y-coordinates are 80x80 meters. In the topologies solid lined circle signifies the CS range of a source node and dotted line circle denotes the CS of receiving node. The transmission range of the source node is represented by pink colored line circle. In Fig. 6 each topology consist of non-coordinated interference links and coordinated interfering links. In Fig. 5, 6, 7, 8 and 9, the carrier sensing range of WMN topologies is reduced gradually. Spectral Gap (Sgap) represents the carrier sensing range of a node in the network denoted as spectrum gap. In Fig. 5, sgap is taken as 0 which illustrates that it is the maximum carrier sensing range of a node. The IA links are identified for that topology. For link (1,2) the identified IA links are (3,4) (12,13) (13,14) and (14,15). The coordinated interfering links for the link (1,2) are (2,3) (3,4) (7,8) (16,17) (17,18) (21,22) (22,23) (23,24). When the carrier sensing range is maximum, the IA interference will be more.

In Fig. 6 the value of Sgap is 1, it represents that the carrier sensing range reduce in sizes and reduced. In this case the IA link for (1,2) is (13,14) and the coordinated interfering links will remain the same. Further in Fig. 7 when sgap equals 2 the carrier sensing range minimizes and the IA links for (1,2) are (13,14). The Fig. 8 illustrates that when the sgap equals 3 than the identified IA links for node 1 and 2 are (3,4) (12,13) respectively. In Fig. 9 the bare minimum

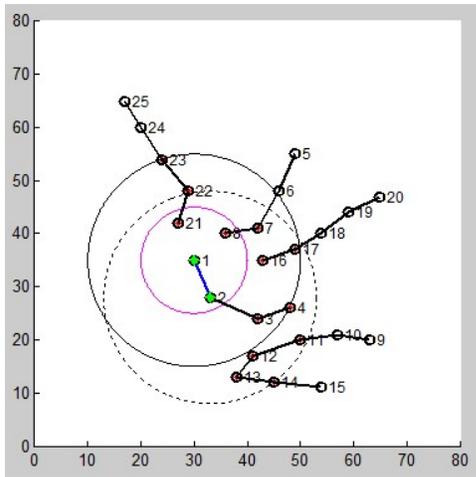


Fig. 5. WMN Topology for spectral gap 0.

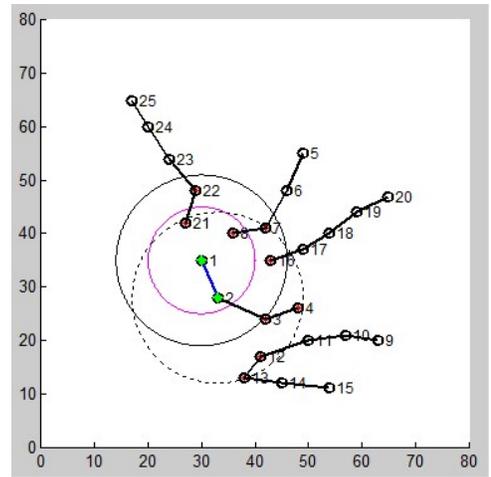


Fig. 7. WMN Topology for spectral gap 2.

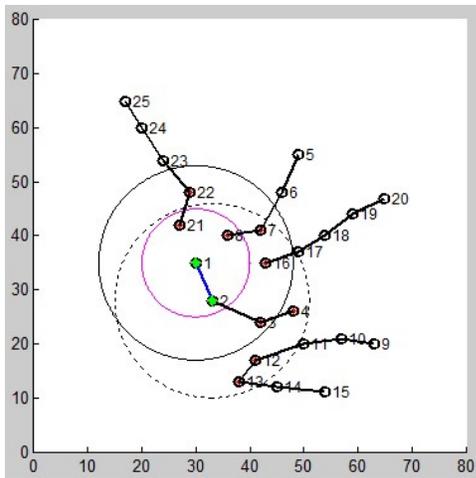


Fig. 6. WMN Topology for spectral gap 1.

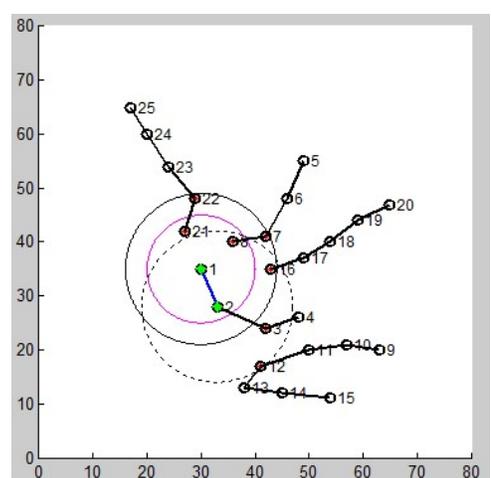


Fig. 8. WMN Topology for spectral gap 3.

carrier sensing range is acquired. Moreover when the value of sgap is equal to 4 the carrier sensing range excessively decreases and the IA interference link is (3,4). This process is repeated for other topologies also. The outcome of information asymmetry interference is greater and the data rate will be highly decreased. In order to minimize IA interference between the nodes optimal channel assignment scheme is used.

Table III displays a list of coordinated interfering (CO) links and IA links. The first column shows the intended links, the second column represents coordinated interfering links. The third column represents Information Asymmetry interfering links. For identifying IA links each time the carrier sensing range is reduced. The sgap starts from value 0 and approaches to 4. The value 0 represents maximum sensing range, the range decreases when the value reaches to 4. The Carrier sensing range is reduced at sgap equals 4. These links are derived from Fig. 5.

Fig. 10 and 11 represents two more topologies comprising of 25 nodes each. The terrain area has been taken 80x80 meters. The nodes alignment is based on x-axis and y-axis. In topology 10 and 11 four paths are taken. The black solid circle represents CS of source node and dotted circle represented

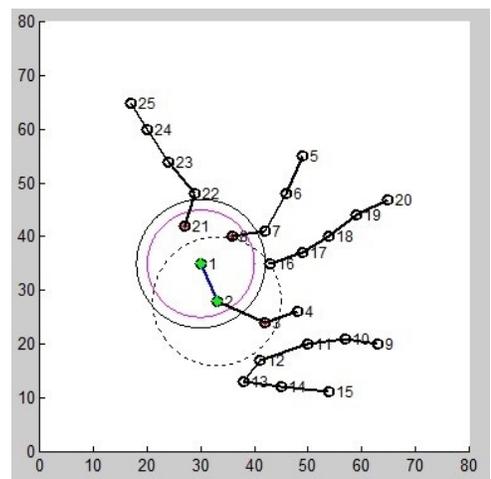


Fig. 9. WMN Topology for spectral gap 4.

CS range of destination node whereas pink circle represented transmission range of source. For both the topologies coordinated links and information asymmetry links have been

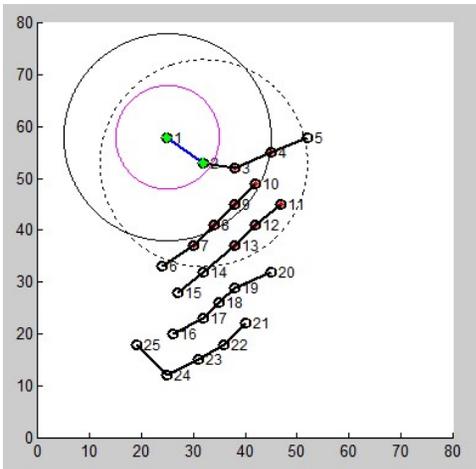


Fig. 10. MPMC-WMN MATLAB generated 25 node topology 1.

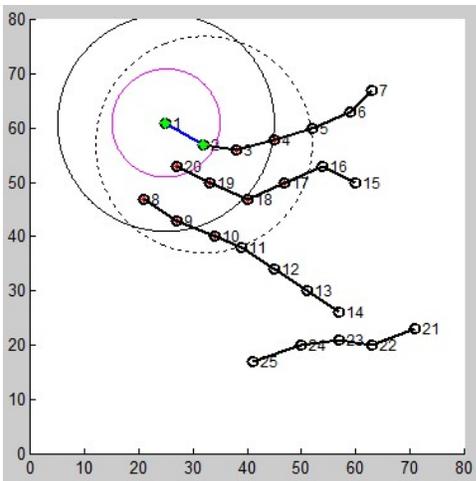


Fig. 11. MPMC-WMN MATLAB generated 25 node topology 2.

identified. The carrier sensing range is decreased by spectral gap value. The value of sgap ranges from 0 to 4. At each sgap value the information asymmetry links are identified.

B. Partially Overlapping Channel (POC) assignment in AMPL

AMPL displays results regarding channel assignment. Partially Overlapping Channels (POCs) are assigned to the links in MPMC wireless mesh network. AMPL assign same traffic load to each link. That is varied from 50 to 500 packets/ sec. The spectrum IEEE 802.11b has 11 channels. The partially overlapping channels are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 have been considered for channel assignment. AMPL utilizes the gurobi solver for channel assignment results. The linear programming models have been executed. The file created is a text file that comprises of real execution of the proposed model. In gurobi solver the solve command gives the output or goal of the proposed model. The binary variable xx displays channel assignment and link. For instance xx [*,*] shows xx[source link, destination link, channel assigned].

The variable xx[2,3,1] illustrates that channel 1 is allotted

TABLE III. LIST OF COORDINATED INTERFERENCE AND IA INTERFERENCE LINKS OF WMN TOPOLOGY

LINK	COORDINATED LINKS	IA LINKS
(1,2)	(2,3)(3,4)(7,8)(16,17)(17,18) (21,22)(22,23)(23,24)	(3,4)(12,13) (14,15) (13,14)
(2,3)	(1,2)(3,4)(7,8)(11,12)(12,13)(13,14) (14,15)(16,17)(17,18)(21,22)	(13,14)(14,15) (16,17)(17,18)
(3,4)	(1,2)(2,3)(7,8)(10,11)(11,12)(12,13) (13,14)(14,15)(16,17)(17,18)	(17,18)(18,19)
(5,6)	(6,7)(7,8)(17,18)(18,19)(19,20)	(7,8)(18,19) (19,20) (22,23)
(6,7)	(5,6)(7,8)(16,17)(17,18)(18,19) (19,20)(21,22)(22,23)	(1,2)(2,3) (22,23)
(7,8)	(1,2)(2,3)(3,4)(5,6)(6,7)(16,17)(17,18) (18,19)(19,20)(21,22)(22,23)	(1,2) (2,3) (22,23) (23,24)
(9,10)	(10,11)(11,12)(14,15)	(11,12) (12,13) (17,18) (18,19)
(10,11)	(3,4)(9,10)(11,12)(12,13)(14,15)(16,17)(17,18)(18,19)	(12,13)
(11,12)	(2,3)(3,4)(9,10)(10,11)(12,13)(13,14)(14,15)(16,17)(17,18)	Nil
(12,13)	(2,3)(3,4)(10,11)(11,12)(13,14)(14,15)(16,17)	Nil
(13,14)	(2,3)(3,4)(11,12)(12,13)(14,15)	Nil
(14,15)	(2,3)(3,4)(9,10)(10,11)(11,12)(12,13)(13,14)	Nil
(16,17)	(1,2)(2,3)(3,4)(6,7)(7,8)(10,11)(11,12)(12,13) (17,18)(18,19)(19,20)(21,22)(22,23)	(19,20) (18,19)
(17,18)	(1,2)(2,3)(3,4)(5,6)(6,7)(7,8)(10,11) (11,12)(16,17) (18,19)(19,20)	(19,20)
(18,19)	(3,4)(5,6)(6,7)(7,8)(10,11)(16,17) (17,18)(19,20)	Nil
(19,20)	(5,6)(6,7)(7,8)(16,17)(17,18)(18,19)	Nil
(21,22)	(1,2)(2,3)(6,7)(7,8)(16,17) (22,23)(23,24)(24,25)	(23,24) (24,25)
(22,23)	(1,2)(6,7)(7,8)(16,17)(21,22)(23,24)(24,25)	(24,25)
(23,24)	(1,2)(7,8)(21,22)(22,23)(24,25)	Nil
(24,25)	(21,22)(22,23)(23,24)	Nil

TABLE IV. OPOCA CHANNEL ASSIGNMENT OF FIGURE 7

LINKS	POC CHANNEL ASSIGNMENT	LINKS	POC CHANNEL ASSIGNMENT
(1,2)	11	(13,14)	6
(2,3)	1	(14,15)	10
(3,4)	6	(16,17)	1
(5,6)	3	(17,18)	1
(6,7)	9	(18,19)	11
(7,8)	8	(19,20)	1
(9,10)	1	(21,22)	1
(10,11)	4	(22,23)	1
(11,12)	6	(23,24)	4
(12,13)	7	(24,25)	1

to a link (2,3). Same as the case with xx[5,6,3] shows that the channel 3 is allotted to link (5,6) in a network. The total numbers of partially overlapping channels are 11, So the value of xx will proceed to 11. Similarly for channel 11 this concept will be used, the variable xx[10,11,4] shows that 4 channel is allocated to the link (10,11) in WMN. The traffic flow is taken as 0.2, 0.4, 0.8, 1.0 and it exceeds to the value 2.0. These outcomes are taken further for all the demands changing from 50 to 500 packet/ sec. Table IV is completely based on Fig. 7. Table IV consists of two columns. The first column comprises considered links and second column consists of POC channles that are assigned to each link.

C. Information Asymmetry Minimization (IAMin) in AMPL

In the next step orthogonal channels have been assigned to the same topology for the purpose of comparison of both non-overlapping and partially overlapping channels. The results of channel assignment for partially overlapping channels and non-overlapping has been compared. Table V shows channel assignment result for orthogonal channel that are executed through AMPL. The entire orthogonal channel 1, 6, 11 have been assigned to the links. The snapshots have been taken

TABLE V. NON-OVERLAPPING CHANNEL ASSIGNMENT FOR IAMIN MODEL

Links	Non-Overlapping Channel Assignment
(1,2)	11
(2,3)	1
(3,4)	6
(5,6)	1
(6,7)	1
(7,8)	11
(9,10)	1
(10,11)	1
(11,12)	1
(12,13)	11
(13,14)	1
(14,15)	1
(16,17)	1
(17,18)	1
(18,19)	11
(19,20)	1
(21,22)	1
(22,23)	1
(23,24)	1
(24,25)	11



Fig. 12. OPNET generated 25 nodes WMN topology.

during the implementation in AMPL software. The binary variables 0 and 1 are used for orthogonal channel assignment. The binary variable 0 represent no channel has been assigned or absence of channel assignment and 1 represents presence of channel assigned to a link. AMPL assign same traffic load to each link. That is varied from 50 to 500 packets/sec.

Table V represents the orthogonal channel assignment to each of the link in a network. The channel 1, 6 and 11 have been assigned to the links.

D. OPNET Simulation Results

The channel assignment result obtained from POC and non-overlapping channel assignment through AMPL is further used for simulation process in OPNET. All the three topologies are recreated in OPNET modeler. A network model is created by selecting the area and the location of various nodes in that area. Fig. 12 shows 25 nodes topology that is created with the help of OPNET modeler. Three different topologies each consisting of 25 nodes are constructed and the effect of Information Asymmetry interference in a network have been analyzed for both POC and orthogonal channel assignment.

1) *OPNET simulation results for POC model:* The channel assignment results obtained from AMPL is used by OPNET simulator for further process. The links have been divided into

TABLE VI. SIMULATION PARAMETERS FOR THE OPOCA MODEL

PARAMETER	VALUE
Radio Technology	IEEE 802.11b
Radio Frequency	2.4 GHz
Data Rate	11Mbps.
Minimum Bandwidth	22 MHz
Buffer size	25600000 bytes
Transmission Range (Tr)	10 meters
Carrier Sensing Range (CSr)	2*10 meters
Number of Nodes	25
Radios per Node	3
Network Simulator used	OPNET
Simulation time	4 minutes
Terrain area	80x80 m2
Total scenarios	3

TABLE VII. PROPOSED OPOCA MODEL SIMULATION RESULTS

Flow (Packets/Sec)	Topology 1	Topology 2	Topology 3	Average
50	924.45	1029.59	958.17	970.736667
100	1720.77	1969.27	1894.46	1861.5
150	1957.53	2689.41	2663.93	2436.95667
200	2061.05	3025.85	3037.48	2708.12667
250	2175.03	3404.41	3177.13	2918.85667
300	2264.11	3685.18	3324.98	3091.42333
350	2355.35	3919.37	3459.72	3244.81333
400	2481.11	4133.76	3571.12	3395.33
450	2577.25	4183.24	3657.67	3472.72
500	2670.51	4329.82	3828.99	3609.77333

coordinated and IA interfering links. The OPNET simulation results are gathered for the proposed model that is Optimized Partial Overlapping Channel Assignment (OPOCA) model. The parameters for simulation are shown in Table VI.

Table VI depicts that the radio technology used for each scenario is IEEE 802.11b spectrum. The frequency used is 2.4 GHz and transmission capacity is 11 Mbps. Each node comprises of three radios and different channels are allotted to each radio interface. The demand of flow ranges from 50 to 500 packets/sec. The total simulation time is four minutes. The results obtained from the simulations shows that the proposed model gives better result as compared to the existing model.

Table VII depicts the simulation results of the effect of IA interference over POC channel assignment. The network capacity of every topology has been measured. The network flow ranges from 50 to 500 (packets/sec) for each topology. The table shows that as the flow demand increases the network capacity value also increases. Minimizing the IA interference increases the overall throughput. The last column shows the average result obtained from different topologies of the OPOCA model. For each topology the network capacity varies depending on the density of a network.

Table VIII represents the results we got from existing IAMin model. The traffic flow varies from 50 to 500 packets per second for each topology. The table shows that as the flow demand increases the network capacity value also increases. Minimizing the IA interference increases the overall throughput. The last column represents average network capacity of the existing model. The transmission capacity of the exiting model is lower than that of the proposed model which clarifies that the OPOCA model works better than IAM model.

TABLE VIII. IAM MODEL SIMULATION RESULTS

Flow (Packets/Sec)	Topology 1	Topology 2	Topology 3	Average
50	979.21	1038.79	858.25	958.75
100	1783.09	1840.58	1694.1	1772.59
150	2239.82	2328.93	2174.34	2247.69667
200	2485.53	2619.65	2580.12	2561.76667
250	2626.37	2853.6	2836.3	2772.09
300	2758.57	3037.53	3009.57	2935.22333
350	2837.99	3144.7	3235.9	3072.86333
400	2935.28	3264.71	3318.54	3172.84333
450	2972.79	3369.85	3446.81	3263.15
500	2982.99	3450.61	3532.84	3322.14667

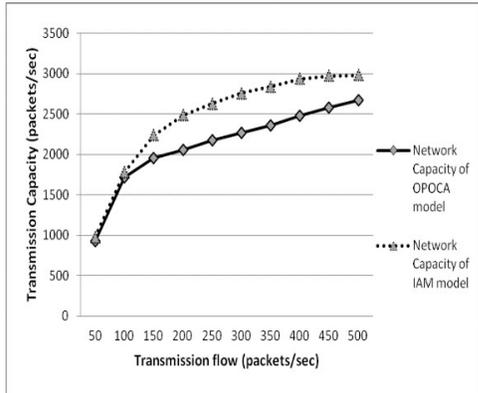


Fig. 13. Capacity comparison of OPOCA model over IAM model for topology 1

Fig. 13 shows the graphical representation of capacity comparison of OPOCA model over the IAMin model. Both the model consists of 25 nodes each. The horizontal axis shows the transmission flow which varies from 50 to 500 packets/sec, whereas the vertical line shows the transmission capacity.

The above dotted line shows the increment in the network capacity of IAMin model and the below solid line represent increase in network capacity of the POC model. From the below graph it has been verified that where the IA interference is minimum the channel assignment of orthogonal channel (IAM) works better than the partially overlapping channel assignment. The network capacity in sparse network of IAM model is maximum than the OPOCA model, where the nodes are aligned far from each other.

Fig. 14 illustrates the capacity comparison of OPOCA model over the existing model. The graph values have been taken from Tables VII and VIII, based Topology 2. Here an x-axis displays sender node demand on every node, whereas y-axis displays the WMN capacity of network in packets/sec. The above solid line represent network capacity of OPOCA model and below dotted line represents the network capacity of IAMin model. As the flow demand increases the network capacity also increases. The peak value for OPOCA model is 4329.82 and the peak value of existing model (IAM) is 3450.61, from which it is verified that partially overlapping channel assignment works better than IAMin model. In dense network where IA interference is more the partially overlapping channel assignment gives better performance.

Fig. 15 indicate the network capacity relationship among the OPOCA and IAMin model for third topology. For the proposed model the throughput value ranges from 958.17 kbps for 50 packets/sec and increases to 3828.99 kbps for 500

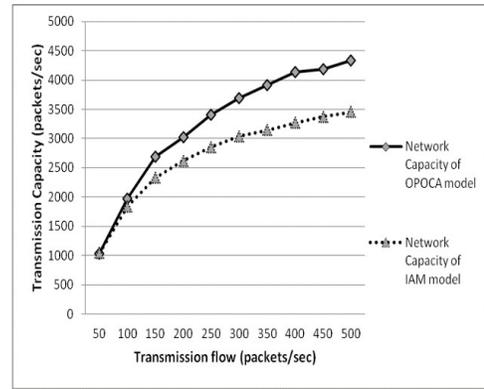


Fig. 14. Capacity comparison of OPOCA model over IAM model for topology 2.

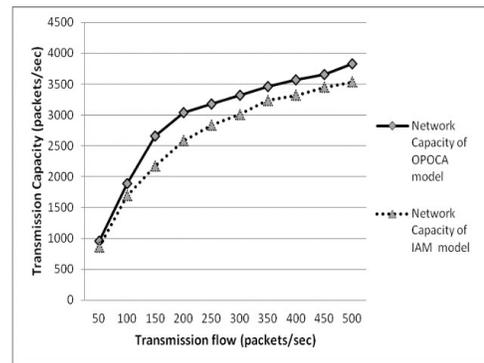


Fig. 15. Capacity comparison of OPOCA model over IAMin model for topology 3.

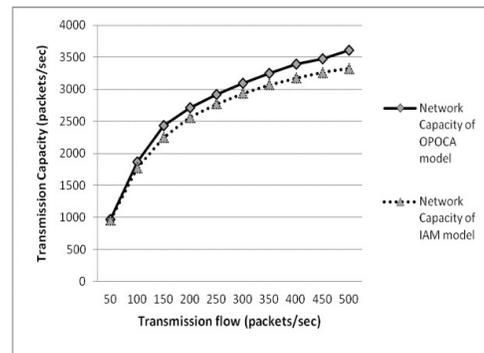


Fig. 16. Capacity comparison of OPOCA model over IAMin model for topology 4.

packets/sec.

As the traffic flow increases capacity also increases. The IAMin model value ranges from 858.25 kbps for 50 packets/sec and increases to 3532.84 kbps for 500 packets/sec. The graphical representation of data shows that the network capacity of OPOCA model is more than that of IAMin model.

Fig. 16 represents average network comparison of capacity between OPOCA model and IAM model. The upper solid line in graph represents average capacity of an OPOCA model and the lower dotted line shows the average capacity of IAMin model. The horizontal value shows traffic demand in packet/

TABLE IX. NETWORK CAPACITY IMPROVEMENT OF OPOCA MODEL OVER IAMIN MODEL

Flow (Packets/Sec)	Avg Network Capacity Of IAMin Model	Avg Network Capacity Of OPOCA Model
50	958.75	970.7367
100	1772.59	1861.5
150	2247.697	2436.957
200	2561.767	2708.127
250	2772.09	2918.857
300	2935.223	3091.423
350	3072.863	3244.813
400	3172.843	3395.33
450	3263.15	3472.72
500	3322.147	3609.773

TABLE X. NETWORK CAPACITY IMPROVEMENT OF OPOCA MODEL OVER IAMIN MODEL

Flow (Packets/Sec)	Percentage Capacity Of IAMin Model	Percentage Capacity Of OPOCA Model	Percentage (%) Improvement
50	95	97	2
100	88	93	5
150	75	81	6
200	64	68	4
250	55	58	3
300	48	51	3
350	43	46	3
400	39	42	3
450	36	38	2
500	33	36	3

sec for the topology of 25 nodes. Similarly the vertical line in the graph shows the normal network capacity for those models i.e. consist of existing and proposed model. The graph shows that the average network capacity is improved by POC model, and gives better performance than the IAMin model. As the IA interference is minimized by optimal channel assignment the overall network capacity increases. Tables IX and X represents that the average network capacity of the proposed model that is 62% and the network capacity of the existing IAM model is 57%. The percentage improvement between existing and proposed model is 5% that shows, the Optimized Partially Overlapping Channel Assignment (OPOCA) model provides better capacity improvement in a dense environment over the Information Asymmetry Minimization (IAMin) model.

V. CONCLUSION

In this research the information asymmetry interference is minimized and WMN network capability is maximized. The OPOCA model is compared with the IAM model that is without the node radio constraint and partially overlapping channel constraint. The simulation has been carried out on 25 node topology and the result shows that the proposed channel assignment scheme gives better result where the information asymmetry is high, than the existing model. The average network capacity of the proposed model is 63% and the network capacity of the existing IAMin model is 55%. The percentage improvement between existing and proposed model is 8% that shows, the Optimized Partially Overlapping Channel Assignment (OPOCA) model provides better capacity improvement in a dense environment over the Information Asymmetry Minimization (IAM) model. Through simulation results it is verified that OPOCA model provides considerable capacity improvement over the IAMin model.

ACKNOWLEDGMENT

The authors would like to thank the University of Agriculture Peshawar and FATA university FR, Kohat Pakistan for supporting the research done in this paper. The co-authors also participated enthusiastically and did their full efforts to explore the channel assignment and its results in Wireless mesh networks.

REFERENCES

- [1] Gledson, E., C. Daniel, P. Ferreira, and Gustavo (2011) A Layered Routing Architecture for Infrastructure Wireless Mesh Networks. In Networking and Services, 2009.ICNS'09. Fifth International Conference on, pp. 366-369.
- [2] Chen, Y., N. Xie, G. Qian, and H. Wang, (2010). Channel assignment schemes in Wireless Mesh Networks, In: Mobile Congress (GMC), 2010 Global , vol., no., pp.1-5, 18-19 Oct. 2010
- [3] Shah, S., H. Hussain, and M. Shoaib, (2013, September). Minimizing non-coordinated interference in multi-radio multi-channel Wireless Mesh Networks (MRMC-WMNs). In Digital Information Management (ICDIM), 2013 Eighth International Conference on, pp. 24-28.IEEE, 2013.
- [4] Shah, S., H. Hussain, and M. Shoaib, (2013, September). Minimizing non-coordinated interference in multi-radio multi-channel Wireless Mesh Networks (MRMC-WMNs). In Digital Information Management (ICDIM), 2013 Eighth International Conference on, pp. 24-28.IEEE, 2013.
- [5] Shinde, S. R., M. L. Dhore, and J. B. Karande, (2009) Avoiding Interferences in WLAN 802.11b for PartiallyOverlapped Channels.International Conference on Advances in Computing, Communication and Control (ICAC309)
- [6] Weisheng, S. (2010) An overview of Channel Assignment methods for multi-radio multi-channel wireless mesh networks Volume 70, Issue 5, Pages 505524.
- [7] Pollak, S. and V. Wieser, (2012) Channel Assignment SchemesOptimization for Multi-InterfaceWireless Mesh Networks Based on Link Load. INTECH Open Access Publisher, 2012.
- [8] Venkata, K. and M. M. Naidu, (November, 2014) Evaluation of Interference-Aware Channel Allocation Algorithms for Wireless Mesh Networks.International Journal of Emerging Research in Management & Technology ISSN: 2278-9359 (Volume-3, Issue-11)
- [9] Wang, J., W. Shi, K. Cui, F. Jin, and Y. Li, (2015) Partially overlapped channel assignment for multi-channel multi-radio wireless mesh networks. EURASIP Journal on Wireless Communications and Networking (2015)
- [10] Garetto, M., T. Salonidis, and E. W. Knightly, (2008). Modeling per-flow throughput and capturing starvation in CSMA multi-hop wireless networks. IEEE/ACM Transactions on Networking (TON), 16(4), 864-877.
- [11] Shah, S., A. W. Abbas, H. Hussain and H. Israr, (2017) Minimizing information asymmetry interference in multi-radio multi-channel wireless mesh networks. Kuwait Journal of Science, 44(3) pp. 30-39.
- [12] Shah, S., M. Atif, S. Khan, M. Daud and F. K. Khalil (2018). A Comparison of Near-Hidden and Information Asymmetry Interference Problems in Wireless Mesh Networks. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 4, 2018

Initialization Method for Communication and Data Sharing in P2P Environment Between Wireless Sensor Nodes

M. Asif Jamal¹

Department of Computer Science
& Engineering,
Air University,
Islamabad, Pakistan

Aziz Ur Rehman²

Department of Computer Science & IT
The University of Lahore,
Gujrat, Pakistan

Moonisa Ahsan³

Department of Computer Science & IT
The University of Lahore,
Gujrat, Pakistan

M. S. Riaz⁴

Department of Computer Science
& Engineering,
Air University,
Islamabad, Pakistan

M. S. Zafar⁵

Department of Computer Science
Beihang University of Aeronautics
& Astronautics (BUAA),
Beijing, China

Abstract—Wireless Sensor Networks have increased noteworthy thought nowadays, rather than wired sensor systems, by presenting multi-useful remote hubs, which are littler in size. However, WSNs correspondence is inclined to negative impacts from the physical environment, like, physical hurdles and interference. The reason for this work is to outline a testbed, to introduce method for communication startup and data sharing in a peer to peer (p2p) environment between wireless sensor nodes. The work is directed on both the IEEE 802.15.4 physical and the application layers. In this testbed, one channel, from the IEEE 802.15.4 channels range is devoted as an “emergency channel” which is utilized for handshaking or in case there is communication failure between the Transmitter (Tx) and Receiver (Rx) nodes. The remaining 15 channels are called “data channels” and are utilized for real information transmission and control signals. Linux based TinyOS-2.x is utilized as a working framework for low power sensors. MICAz bits are utilized as nodes and a MIB520 programming board is utilized for burning the codes and for the purpose of gateways.

Keywords—TinyOS; peer-to-peer; motes; testbed; nesC; MICAz; MIB520; handshaking

I. INTRODUCTION

An awesome improvement is going on these days with the generation of low power remote sensors. However, during the communication of these low power enabled sensor nodes, interferences must be avoided from any source internally or externally. Therefore, it is necessary to make a testbed to evaluate result in an environment that eliminates the internal or external source of interference. Testbed for wireless sensor networks can play an important role in academia because, theoretical study and simulation show results in ideal situation. A wireless sensor network consists of low power sensor nodes, which have the responsibility to sense the task assigned to them and report this sensed information via some wireless link to gateway. Usually these nodes comprise of a microchip, which is responsible for transmission and reception of data [1], [2].

In WSNs a good testbed ought to have these properties,

- 1) Maintain synchronization in the occurrence of communication failure.
- 2) Should have the ability of P2P communication symmetry.

Section 2 presents literature review and background study whereas Section 3 described the methodology followed by Section 4 that includes implementation and results. Section 5 shows conclusion and future work.

II. LITERATURE REVIEW

This work is planned to make a testbed which conveys improvement to a current testbed created at Mid-Sweden University (MIUN) “An Empirical Study of Low Power Multichannel correspondence in WSN”, created by authors [3], which, over the long haul, will be advantageous for outlining new conventions. On account of correspondence misfortune, a calculation is intended for the synchronization of nodes. For transmitter-receiving symmetry the calculation is outlined so that after a particular number of packets have been sent by the transmitter, both nodes change their part.

The correspondence depends on the IEEE 802.15.4. It uses carrier sense multiple access with collision avoidance (CSMA/CA) as an access provision method. As CSMA/CA works on low data application, so it provides enough throughput without severe interference and delay [4]. This IEEE 802.15.4 gives an aggregate 26 channels [5] and from this, one lies in the 868 MHz band (utilized as a part of Europe later extended to three in 2006 [6]), 10 lies in the 915MHz band (utilized in North America extended to thirty in 2006 and 16 lies in the 2.4GHz band (utilized around the world). The 2.4GHz variant of IEEE 802.15.4 offers the most astounding throughput of 250 kbps, and is, hence, covers long distance [7]. Subsequently this testbed concentrates on the 2.4 GHz

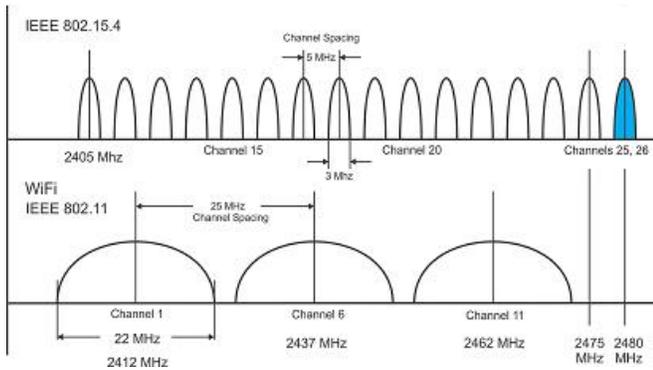


Fig. 1. IEEE 802.15.4 and Wi-Fi Spectrum [8].

range of 16 channels ranging from channel no 11 to 26. Fig. 1 shows the IEEE 802.15.4 and IEEE 802.11 Spectrum.

The existing MIUN testbed [3], for WSN has shown problems. The main issue was the correspondence failure between the gateway and the nodes for longer interval of time, which results loss of important time during the field trial. The objective of this work has been to respond to the following questions.

- 1) How to synchronize the nodes when the communication is lost due to hardware problem or inter-channel interference.
- 2) Implementation of peer to peer communication between transmitter and receiver nodes.

A. Definition and Implementation of a New Initialization Algorithm

In previous studies [3], after five unsuccessful beacon message transmissions, the transmitter node (Tx) switches to the next channel, and the Receiver Node (Rx) will remain on the current channel as it is not receiving beacon, meaning that it must wait until the Tx has made a full sweep over all channels, for a new communication attempt to occur. The new algorithm should eliminate this problem. The algorithm can be in the form of a handshake or any other method which is robust to packet loss. The new algorithm should ensure that the effort to establish the communication will be made more efficient.

B. Handshaking

In communication, handshaking mechanism between nodes is considered to be most significant for connection establishment. Whenever we wish to establish or re-establish a connection, handshaking is the foremost step. Therefore, handshaking can be defined as the process in which Tx broadcasts a number of beacons, which, if the signal is received efficiently to the Receiver node (Rx), then the Rx will response with an acknowledgment (ACK) message. The ACK message shows the successful agreement between sender and receiver nodes for conducting efficient handshaking for peer-to-peer communication [9]. The same technique will be implemented in this work and one channel from the IEEE 802.15.4 spectrum (channel # 26) with proficiency of 2.4 GHz bandwidth will be dedicated to establish hand-shaking path between nodes.

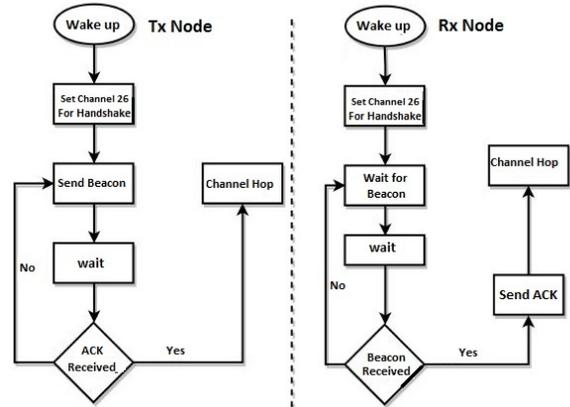


Fig. 2. Flowchart for Tx & Rx Hand Shake.

C. Introducing Peer to Peer (P2P) Symmetry between the Transmitter-receiver Pair

In the previous setup [3], the Tx triggers the communication with Rx via beacons. After sending 5 beacons Tx starts the transmission of data packets, while the Rx stores the packet logs. In our proposed setup, the code should be modified so that, in the first round after a predefined number of packets transmitted by Tx, the two communicating nodes should switch roles and the direction of communication changes. In other words, the node that has been receiving and logging data should, in the following round, take up the transmitter role, and the former transmitter should maintain packet logs. This task also comprises the introduction of acknowledgment packets.

D. Channel Hopping/Frequency Hopping

Channel Hopping is the consecutive change of channel in the available frequency range. Because of obstruction in the remote medium [10], channel hopping must be performed in a manner that both Tx and Rx might change channels at the same time. In the testbed channel hopping calculation must be characterized for both Tx and Rx.

III. METHODOLOGY

In the IEEE 802.15.4 range, channel 26, is termed as an “Emergency Channel” in this testbed, while the remaining channels (11 to 25) are called “Data Channels”. At first, for peer to peer correspondence, it is vital that the two nodes must handshake with each other before an information trade can happen. In this way, the emergency channel is dedicated for handshaking between the Tx and Rx. The data communication channels are devoted for actual information transmission. Fig. 2 demonstrates the handshaking process.

In Fig. 2 when the Tx is started transmission at the emergency channel it then begins sending beacons occasionally and waits for an affirmation in the form of ACK from Rx. If Rx receives the beacon, then it replies with an acknowledgment and jumps to the first data channels (e.g. Channel 11). Tx likewise hops to the same data channel, predefined to both Tx and Rx.

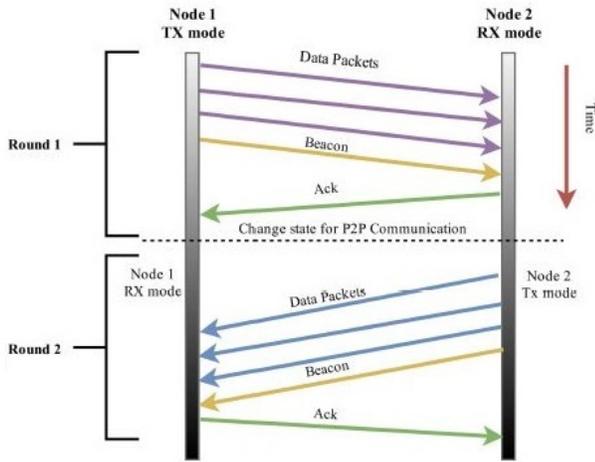


Fig. 3. P2P data transmission symmetry in a single data channel.

Keeping in mind the end goal to acquaint the symmetry with the P2P pair at the data communication channels, a packet counter is utilized to record the number of packets transmitted by the Tx. When the maximum packet counter limit is reached, a beacon is sent by Tx to remind Rx, to change its transmission direction (e.g. from Rx to Tx). Upon the successful reception of beacon message by Rx, it begins sending data packets to Tx to form a peer to peer communication. At the point when the maximum packet counter point is reached, again a beacon is sent by Rx and it waits for an ACK from Tx. After receiving the beacon by Tx, it sends an ACK to Rx, and hops to the next channel. At this point, successful reception of ACK, Rx likewise hops to the following data channel. The above methodology is reshaped in all data channels. Fig. 3 demonstrates a diagrammatic representation of the Tx and RX P2P transmission symmetry. If a packet/channel loss or hardware failure occurs, then the nodes likewise hop back to the emergency channel for handshaking.

For explaining the Task 2 more deeply, flow-chart diagrams and following terms will be considered.

At the emergency channel:

- 1) The Tx is capable of sending beacon only to check the communication path with Rx for hand-shaking process.
- 2) The ACK is only send by the Rx after the successful receiving of the beacon packet.

At the data channel:

- 1) After the last data packet (i.e. the final packet preceded by the beacon), beacon is sent to the destination. If the beacon is successfully received by Rx, it shows that a shift of the transmission roles should occur, and if the beacon is received by Tx, it shows that it should hop to the next available data channel in network.
- 2) ACK is sent upon successful reception of the Beacon.

Peer-to-peer communication among two motes is implemented in such a way that the communication path will alter whenever the beacon packet is sent by Tx, after sending a

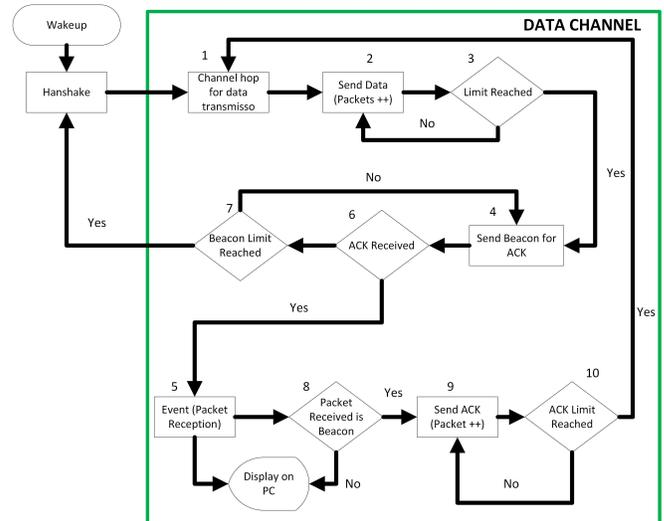


Fig. 4. Tx flow chart for p2p communication.

specified number of data packets and is successfully received by the Rx Channel. Fig. 4, the flow-chart diagram of Tx Peer-to-peer communication is illustrated. When the Tx hops from emergency channel to the initial channels by means of the channel hopping task (1), the Tx initiates the process of sending data packets (2). When the defined maximum packet limit is reached, Tx then halts the process of sending further data packets (3) and sends a beacon (4) to alert Rx to take its turn of communication. If a successful ACK is received (6) from Rx then the Tx node gets into waiting mode for receiving data packets from Rx (5). During this process, while Tx is receiving data packets, if a beacon is received (8) from Rx then Tx responds back with an ACK Alert (9) (according to our testbed criteria, it will send 3 ACK (10)) and then hops to the next available data channel in the network by altering its path using channel hopping task (1).

A. Hardware Failure or Communication Loss

In Fig. 4 if Tx is not receiving an ACK (6) from Rx, it means that communication loss or hardware failure has occurred, then Tx will keep sending beacons until it reaches its maximum beacon limit (maximum beacon limit = 3). If no ACK is received by Tx based on the last beacon sent (7), then it will jump to the emergency channel.

B. RX Flowchart

In Fig. 5 below the flow charts inside green border shows RX P2P communication at the data channel where the Rx receives packets from Tx. Every packet received from Tx is forwarded to a serial port for PC logging. If, during the packet receiving, a beacon is received (3) then Rx will reply with an ACK (4) and starts sending data packets (5) until the maximum number of data packet limit, is reached and a beacon is sent (6), to inform Tx that it is time to jump to the next data channel and wait for an ACK from Tx. If an ACK is received (7) then Rx jumps to the next data channel by means of the channel hopping task (1).


```

TX interrupted
nap@NAP: ~/Desktop/opt/tinyos-2.1.0/apps/t
File Edit View Terminal Help
21 65 136 37 34 0 1 0 0 0 63 6 0 0 0 0 0 22 0 0 6 236 crc is 1
20 65 136 38 34 0 1 0 0 0 63 6 0 0 0 0 0 23 0 4 236 crc is 1
23 65 136 39 34 0 1 0 0 0 63 6 0 0 0 0 0 23 0 0 0 5 236 crc is 1
23 65 136 40 34 0 1 0 0 0 63 6 0 0 0 0 0 23 0 0 0 5 236 crc is 1
23 65 136 41 34 0 1 0 0 0 63 6 0 0 0 0 0 23 0 0 0 6 236 crc is 1
21 65 136 42 34 0 1 0 0 0 63 6 0 0 0 0 0 23 0 0 5 236 crc is 1
20 65 136 0 34 0 1 0 0 0 63 6 0 0 0 0 0 26 0 6 234 crc is 1
20 65 136 2 34 0 1 0 0 0 63 6 0 0 0 0 0 11 0 8 231 crc is 1
23 65 136 3 34 0 1 0 0 0 63 6 0 0 0 0 0 11 0 0 0 8 235 crc is 1
23 65 136 4 34 0 1 0 0 0 63 6 0 0 0 0 1 0 11 0 0 0 8 234 crc is 1
23 65 136 5 34 0 1 0 0 0 63 6 0 0 0 0 2 0 11 0 0 0 8 233 crc is 1
21 65 136 6 34 0 1 0 0 0 63 6 0 0 0 0 0 11 0 0 8 235 crc is 1
20 65 136 7 34 0 1 0 0 0 63 6 0 0 0 0 0 12 0 8 235 crc is 1
23 65 136 8 34 0 1 0 0 0 63 6 0 0 0 0 0 12 0 0 0 9 236 crc is 1
23 65 136 9 34 0 1 0 0 0 63 6 0 0 0 0 1 0 12 0 0 0 8 236 crc is 1
23 65 136 10 34 0 1 0 0 0 63 6 0 0 0 0 2 0 12 0 0 0 8 236 crc is 1
21 65 136 11 34 0 1 0 0 0 63 6 0 0 0 0 0 12 0 0 8 235 crc is 1
20 65 136 12 34 0 1 0 0 0 63 6 0 0 0 0 0 13 0 6 234 crc is 1
    
```

Fig. 9. Tx interrupted at channel 23.

channels (11 to 13).

Fig. 10 shows the Rx output in the case of a communication loss with Tx at channel 23 as discussed above. The red square shows the point at which the communication loss has occurred. It is cleared from the beacon packet inside the red border that, due to communication loss, 23 beacons are lost which were sent by Tx, while the handshaking is conducted on the 24th beacon.

A. Performance Evaluation

To evaluate the performance of the existing MIUN testbed [3] with the testbed developed in this work the communication reestablishment time in the case of communication loss will be calculated. It is assumed that the data packets transmitted in each channel are 10,000 the time interval between two consecutive packets is 100ms and the total number of channels are 16. At the time when both Tx and Rx were jumping to channel 11, the communication loss occurs due to the reset button of Rx mote being pressed. When the reset button is pressed, the Rx will jump to the default channel, which is channel 26.

In the MIUN testbed [3], Tx will go through all the

```

nap@NAP: ~/Desktop/opt/tinyos-2.1.0/apps/testis
File Edit View Terminal Help
20 65 136 38 34 0 0 0 1 0 63 6 0 0 0 0 0 22 1 5 236 crc is 1
23 65 136 39 34 0 0 0 1 0 63 6 0 0 0 0 0 23 1 0 0 0 4 236 crc is 1
23 65 136 40 34 0 0 0 1 0 63 6 0 0 0 0 1 0 23 1 0 0 0 4 235 crc is 1
23 65 136 41 34 0 0 0 1 0 63 6 0 0 0 0 2 0 23 1 0 0 0 4 234 crc is 1
23 65 136 42 34 0 0 0 1 0 63 6 0 0 0 0 3 0 23 1 0 0 0 4 236 crc is 1
23 65 136 43 34 0 0 0 1 0 63 6 0 0 0 0 4 0 23 1 0 0 0 4 234 crc is 1
21 65 136 44 34 0 0 0 1 0 63 6 0 0 0 0 0 23 1 0 4 235 crc is 1
20 65 136 45 34 0 0 0 1 0 63 6 0 0 0 0 0 23 1 4 235 crc is 1
21 65 136 24 34 0 0 0 1 0 63 6 0 0 0 0 24 0 26 1 0 5 236 crc is 1
23 65 136 25 34 0 0 0 1 0 63 6 0 0 0 0 0 11 1 0 0 0 7 236 crc is 1
23 65 136 26 34 0 0 0 1 0 63 6 0 0 0 0 1 0 11 1 0 0 0 7 233 crc is 1
23 65 136 27 34 0 0 0 1 0 63 6 0 0 0 0 2 0 11 1 0 0 0 8 235 crc is 1
23 65 136 28 34 0 0 0 1 0 63 6 0 0 0 0 3 0 11 1 0 0 0 8 235 crc is 1
23 65 136 29 34 0 0 0 1 0 63 6 0 0 0 0 4 0 11 1 0 0 0 7 236 crc is 1
21 65 136 30 34 0 0 0 1 0 63 6 0 0 0 0 0 11 1 0 8 236 crc is 1
20 65 136 31 34 0 0 0 1 0 63 6 0 0 0 0 0 11 1 7 236 crc is 1
23 65 136 32 34 0 0 0 1 0 63 6 0 0 0 0 0 12 1 0 0 0 7 234 crc is 1
23 65 136 33 34 0 0 0 1 0 63 6 0 0 0 0 1 0 12 1 0 0 0 7 233 crc is 1
23 65 136 34 34 0 0 0 1 0 63 6 0 0 0 0 2 0 12 1 0 0 0 6 236 crc is 1
23 65 136 35 34 0 0 0 1 0 63 6 0 0 0 0 3 0 12 1 0 0 0 7 236 crc is 1
23 65 136 36 34 0 0 0 1 0 63 6 0 0 0 0 4 0 12 1 0 0 0 7 235 crc is 1
    
```

Fig. 10. Rx output in case of Communication Loss.

TABLE I. PERFORMANCE EVALUATION

No. of Packets	Re-establishment Time(ms)		Packet loss	
	MIUN testbed	Testbed (modified)	MIUN testbed	Testbed (modified)
10	$14 * 10^3$	800	140	5
100	$14 * 10^4$	$5.3 * 10^3$	1400	50
1000	$14 * 10^5$	$5.03 * 10^4$	14000	500
10000	$14 * 10^6$	$5.003 * 10^5$	140000	5000
100000	$14 * 10^7$	$5.0 * 10^6$	1400000	50000

channels and will send packets without knowing whether or not the packet has been received by Rx. Tx will jump to next channels until it reaches at channel 26. Hence, the maximum time for communication reestablishment would be,

$$\text{Time} = (\text{No. of packets sent by Tx in each channel}) * (\text{time interval between two packets}) * (\text{No. of channel Tx will jump})$$

$$\text{Time} = 10,000 \times 100\text{milliseconds} \times 14$$

$$\text{Time} = 1, 40, 00,000 \text{ milliseconds}$$

$$\text{Time} = 14,000 \text{ sec}$$

Now to calculate the reestablishment time taken by modified testbed, reviewing Fig. 3, the data packets sent in each round would be 5,000 plus 3 beacons. When a communication loss occurs, the Tx will know this after sending the 3rd beacons at channel 12 and will jump to the Emergency channel (channel 26). Hence, the maximum time for communication reestablishment will be,

$$\text{Time} = (\text{No. of packets sent by Tx in each channel}) * (\text{Time interval between two packets}) * (\text{No. of channel Tx will jump}) + (\text{No. of beacons sent by Tx in each channel}) * (\text{Time interval between two beacons})$$

$$\text{Time} = (5,000 \times 100\text{milliseconds} \times 1) + (3 \times 100\text{milliseconds})$$

$$\text{Time} = 5, 00,300\text{milliseconds}$$

$$\text{Time} = 500.3 \text{ sec}$$

Table I below shows the performance evaluation of the MIUN testbed [3] and its modified version (developed in this work). The evaluation shows that the communication reestablishment time for the modified MIUN testbed is 28 times less than the existing MIUN testbed [3] i.e. a 28 times faster recovery in the case of communication loss.

V. CONCLUSION AND FUTURE WORK

The primary center in this work, as of now talked about in beginning, was in designing a testbed that makes it possible to study the channel properties during the communication of motes at IEEE 802.15.4 in the 2.4GHz band. A new initialization procedure is implemented at communication start up. The initialization algorithm assists in producing a 28 fold increase in time for the communication re-establishment than is possible when using the existing MIUN testbed, in the case of communication loss between the wireless sensor nodes. To study the link symmetry, P2P communication is achieved between Tx and Rx by introducing a beacon and ACK pair. In the case of communication/hardware failure at any data channel, both Tx and Rx will switch to the emergency channel (IEEE 802.15.4 channel # 26) for handshaking.

The future work could be suggested as to reserve 2 channels (both 25 and 26) instead of only 1 channel (26) for the emergency channel, to provide a backup for the emergency channel. Secondly, P2P communication can be performed using timers instead of packet counter.

REFERENCES

- [1] Prabhu, Boselin, M. Pradeep, and E. Gajendran. "Military Applications of Wireless Sensor Network System." (2017).
- [2] Rao, Raghu Mysore, et al. "Multi-antenna testbeds for research and education in wireless communications." *IEEE Communications Magazine* 42.12 (2004): 72-81.
- [3] Liu, Zuguang. "An empirical study of low power multi-channel communication in WSN Zuguang Liu." (2011).
- [4] Caione, Carlo, Davide Brunelli, and Luca Benini. "Distributed compressive sampling for lifetime optimization in dense wireless sensor networks." *IEEE Transactions on Industrial Informatics* 8.1 (2012): 30-40.
- [5] Zheng, Jianliang, and Myung J. Lee. "A comprehensive performance study of IEEE 802.15. 4." *Sensor network operations* 4 (2006): 218-237.
- [6] Baronti, Paolo, et al. "Wireless sensor networks: A survey on the state of the art and the 802.15. 4 and ZigBee standards." *Computer communications* 30.7 (2007): 1655-1695.
- [7] Del Prete, Massimo, et al. "A 2.4 GHz-868 MHz dual-band wake-up radio for wireless sensor network and IoT." 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2015.
- [8] Balodhi, Meenu, Vishwanath Bijalwan, and Banit Negi. "Zigbee & IEEE 802.11 b (WLAN) coexistence in ubiquitous network environment." arXiv preprint arXiv:1407.0462 (2014).
- [9] Mzid, Rania, et al. "Adapting TLS handshake protocol for heterogeneous IP-based WSN using identity based cryptography." *Communication in Wireless Environments and Ubiquitous Systems: New Challenges (ICWUS)*, 2010 International Conference on. IEEE, 2010.
- [10] Watteyne, Thomas, et al. "Mitigating multipath fading through channel hopping in wireless sensor networks." *Communications (ICC)*, 2010 IEEE International Conference on. IEEE, 2010.
- [11] Kodali, Ravi Kishore, and Narasimha Sarma. "Test bed for wireless sensor networks using XMesh networking protocol." *Advances in Computing, Communications and Informatics (ICACCI)*, 2013 International Conference on. IEEE, 2013.
- [12] Steenkamp, Leon du T., Shaun Kaplan, and Richardt H. Wilkinson. "Wireless sensor network gateway." *AFRICON*, 2009. AFRICON'09.. IEEE, 2009.
- [13] Gay, David, et al. "The nesC language: A holistic approach to networked embedded systems." *Acm Sigplan Notices* 49.4 (2014): 41-51.
- [14] LAN/MAN Standards Committee. "Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs)." *IEEE Computer Society* (2003).
- [15] IEEE P802. 15.4 e-2012. "Part 15.4: Low-Rate Wireless Personal Area Networks (WPANs), Amendment 1: MAC sub-layer."

Formal Specification of Memory Coherence Protocol

Jahanzaib Khan, Muhammad Atif,
Muhammad Khurram Zahoor Bajwa, Muhammad Sohaib Mahmood
Department of Computer Science
and Information Technology
The University of Lahore

Sobia Usman
Department of Computer Science
COMSATS University Islamabad
Lahore Campus

Abstract—Memory coherence is the most fundamental requirement in a shared virtual memory system where there are concurrent as well as loosely coupled processes. These processes can demand a page for reading or writing. The memory is called coherent if the last update in a page remains constant for each process until the owner of that page does not change it. The ownership is transferred to a process interested to update that page. In [Kai LI, and Paul Hudak. Memory Coherence in Shared Virtual Memory Systems, 1986. Proc. of Fifth Annual ACM Symposium on Principles of Distributed Computing.], algorithms ensuring memory coherence are given. We formally specify these protocols and report the improvements through formal analysis. The protocols are specified in UPPAAL, i.e., a tool for modeling, validation and verification of real-time systems.

Keywords—Memory coherence; formal specification; shared memory; address space; analysis

I. INTRODUCTION

In a loosely coupled multiprocessors system, virtual memory is useful due to its parallel infrastructure instead of using memory hierarchy. Application programs can use the shared virtual memory just as they do the traditional virtual memory. The data can be naturally migrated between processors on demands because the shared virtual memory discussed in [1] is not only pages data between physical memory and disk but it is also pages data between physical memory and individual processors, as shown in Fig. 1. The shared virtual discussed in [1] provides address space which is shared among all processors in the loosely coupled distributed memory multiprocessors systems. As the shared virtual memory on the loosely coupled multiprocessors has no physically shared memory and the communication cost between processors is nontrivial. Thus the conflicts are not likely to be solved with negligible delay [2]. The problem that Kai Li faced in building the shared virtual memory was memory coherence problem and in [3] Kai Li et al. are focusing on memory coherence problem for shared virtual memory and they provide a number of algorithms as a to solve memory coherence problem. These algorithms include the Central Manager algorithm in which the manager is just like a monitor. The second Algorithm is Improved Central Manager Algorithm. The detail of these algorithms is provided below. We investigate the algorithms with respect to their functional requirements. Our approach for formal verification is based on model-checking. We formally specify the algorithms using UPPAAL. This is comprehensive analysis of the algorithms provided in [1] along with the verification of detailed functional requirements. We give the formal specification of algorithms in functionalism: the timed automata language of UPPAAL [4].

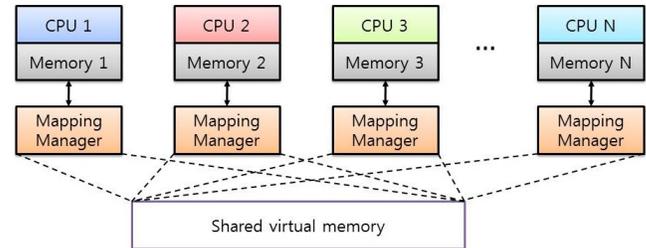


Fig. 1. Shared virtual memory mapping [3].

A. Memory Coherence

A coherent memory means the value returned by a read process is always the same as the value is written by the most recent write process to the same address. In [3] the algorithms for memory coherence are presented. Two of them are:

- 1) Centralized Manger Algorithm.
- 2) Improved Centralized Manager Algorithm.

Each algorithm has the following four basic components:

- **Read Server:** It provides a page as read only.
- **Write Server:** It provides a page for writing.
- **Participant:** A process that is owner of some pages or demands pages for reading/writing.
- **Centralized manager:** It keeps record of all the pages, like who is owner of what and who has taken a page for reading/writing. It is also responsible for changing ownership of a page.

B. Centralized Manager Algorithm

The Central Manager Algorithms maintains a table called info table having tree fields.

- 1) The Owner field contains the processor that is owner of the page and it is the processor which has most recently performed the writ operation on that page.
- 2) The copy-set field contains the list of processors having the copy of the page.
- 3) The Lock field to synchronize the operation.

Each process has also a table called PTable having the fields: access and lock [5]. This table keeps the information about the accessibility of the page on the local processor. In this algorithm there is no fixed owner of the page because

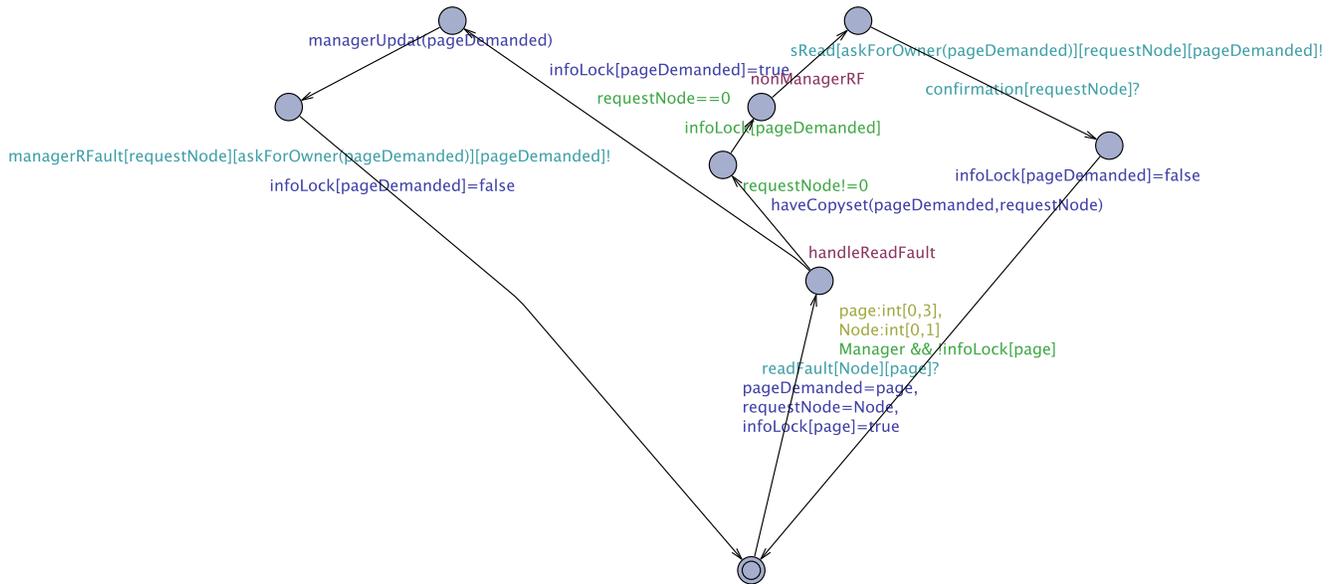


Fig. 3. Read Server Process.

- 6) Process is checked whether it is owner of the page (Boolean).
- 7) All the copy-sets of the page are invalidated in case of write fault.
- 8) Confirmation of successful operation is sent to the manager.
- 9) At completion of operation page is unlocked at manager as well as at owner end.

B. Channels

Synchronization of process is done using channels. Channels receive messages from one process and deliver these messages to other process. There are many types of channels used for synchronization of communication which includes:

- 1) Unicast channel: This channel is mostly used in handshaking of the processes.
- 2) Multicast channel: This channel is used for broadcasting a request in system.
- 3) Urgent channel: At urgent location time progress may not be made. At this location interleaving with normal states is not allowed
- 4) Committed channel: At committed state the possible transition is only one going out of committed state and the committed state has to be left immediately.

How these channels are working in our model is explained below.

C. Central manager algorithm

- 1) readFault [4][pages] is a channel used to send request by client as sender and received by Server where 4 is total number process on which fault can occurred.
- 2) managerRFault [4][4][totalPages] is a channel which is triggered by the manager if and only if the faulting process is manager itself. Manager has information about all pages and the owner of the pages there for manager will send the request to the owner of the page directly and page number is also sent along with.
- 3) sRead[4][4][totalPages] is a channel used to send request to owner of the page by the manager if read fault occurs on non-manager node. It is sending the requested node address and the page demanded by that node.
- 4) serveRead[4][totalPages] is a channel triggered by the owner of the page to the requesting process insure that access for read is granted.
- 5) confirmation [4] is a channel used to send confirmation to the manager for the completion of operation the id of the requesting process is also sent with confirmation.
- 6) writeFault[4][totalPages] is a channel used to send request by client to the server to grant access for writing in the page and if access is granted the requesting process is also declared temporarily owner of the page. On the completion of this operation all the copy-sets of the page are invalidated by the manager of the system.

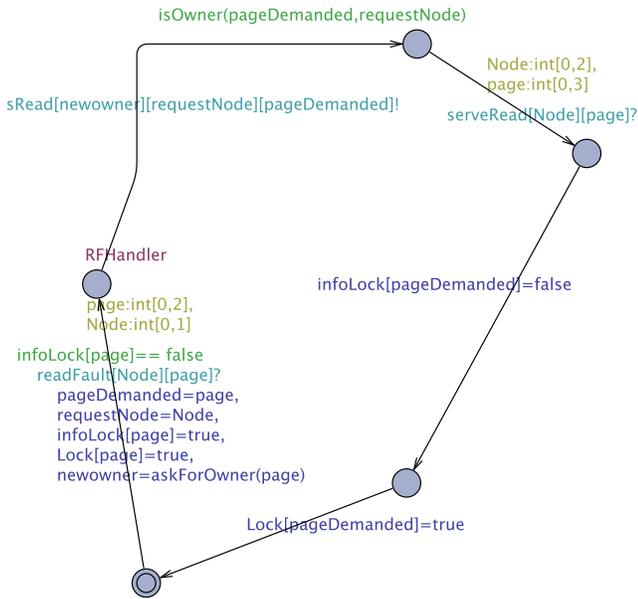


Fig. 4. Improved Read Server Process.

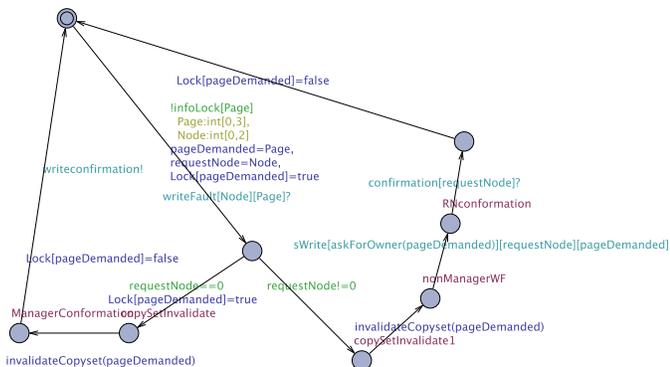


Fig. 5. Write Server Process.

- 7) `sWrite[4][4][totalPages]` is a channel used to send request to owner of the page by the manager if write fault occurs on non manager node. In this channel address of requested node and the page number is also sent as parameters.
- 8) `serveWrite[4][totalPages]` is a channel used by the owner of the page for sending the paged to the requested node for writing. After sending the page the lock of page is removed both by owner of the page and manager of the system and now it is available for new operation.
- 9) `writeConfirmation[totalPages]` is channel used to send the confirmation to the manager of system by the requesting node to assure that it has been granted the access for writing the page.

D. Declarations

Global declarations are made to access the variables throughout the system. The global variables used in our model are placed here.

`const int totalPages=4;` This describes the total number pages for which read or write request can be generated by the processes.

`chan readFault[4][totalPages]` and `chan writeFault[4][totalPages]` are unicast channels and are used to send request to the manager of the system to grant the access for their request.

`chan serveWrite[4][totalPages]`, channel is triggered by the owner of the page to send the page for writing purposes to the requesting node.

`serveRead[4][totalPages]`, channel is used to send the requested page to the requesting node.

`sRead[4][4][totalPages]`, channel is used to send the request to the owner of the page to send the page for reading.

`sWrite[4][4][totalPages]`, channel is used to send the page to the requesting node for writing by the owner of the page.

`chan confirmation[4]`, channel is used to send confirmation of receiving a page to the manager.

`managerRFault[4][4][totalPages]`, channel is used when read fault is occurred on the node which is also manager of the system.

`serveReadManger[totalPages]`, channel is triggered by the manager to request the owner of the page to send the required page to the manager.

`writeConfirmation`, channel is used to send confirmation message to the manager for write fault.

`bool Lock[8]` describes the Lock that is used to lock the page. Lock is made true before serving the page to the requesting process and it is made false when page is served by the owner of the page.

`bool infoLock[8]`, the information of the page is on manager end and before the request is sent to the owner of the page for reading or writing the page the information of the page is locked by making the `infoLock` true and after the it is made

false by the manager after receiving the confirmation message by the requesting process.

int loop,loop2, are used in *invalidateCopySet* function.

```
int [-1,7] pTable[8][8]={  
0,1,1,-1,1,-1,-1,0,  
0,1,1,-1,1,-1,-1,0,  
1,1,1,-1,1,-1,-1,0,  
1,1,1,-1,1,-1,-1,0,  
2,1,1,-1,1,-1,-1,0,  
2,1,1,-1,1,-1,-1,0,  
3,1,1,-1,1,-1,-1,0,  
3,1,1,-1,1,-1,-1,0 }
```

This array contains page table information maintained by the owner of the page. Each row in the array is representing the page number i.e. row number 1 is showing the information about the page number 1 and row number 2 is showing the information about page number 2 and so on for all the other rows in the array. The columns of the array are showing the attributes of the pages which are explained as under.

First column is representing the owner of the page i.e. in row number 1 we have written the 0 in first column which means that the owner of the first page is 0 process and in 7th row there is 3 in first column which shows that the owner of 7th page is process number 3. It is hard coded because in the protocol the ownership is not changed it is only shifted temporarily in write fault case. Second column in the table is representing the information about the access of the page it can be either read or write if the access of the page is marked as read it is made 1 and for write it is made 2 and if the page has both read and write access it is then made 3. Third column in the table is representing the lock information of the page lock. If it is 0 then it means that page is locked and if it is 1 it shows that page is not locked it is unlock and available for operation. Column number 4 to the column number 7 is representing the information about the process having the copy-set of the page i.e. 4th is for first process and 5th is for second if it is 1 against any process this shows that the specific process has the copy-set for that page and if it is 0 it shows that the process has not copy-set of that page. Column number 8 is representing information about either manager has copy-set of the page or not. If it is 1 it shows the manager has the copy-set of that page and if it is 0 it shows that manager has not the copy-set of the process.

```
int [-1,7] copyset[8][5]={  
1,1,-1,-1,0,  
-1,-1,-1,-1,0,  
-1,-1,-1,-1,0,  
-1,1,-1,-1,0,  
-1,-1,-1,-1,0,  
-1,-1,-1,-1,0,  
-1,-1,-1,-1,0,  
-1,-1,-1,-1,0 }
```

This array represents the copy-set table. In this table each row is representing the page number i.e. row number 1 represents page number 1 and row number 2 is representing page number 2 and so on for all rows in the table and columns are representing the other properties of the pages which are as follow.

Columns are representing the processes having the copy of the page. Value 1 means that a particular process has the copy-set of the page and it is maintained in a sequence such as if in row 1 at first column the value is 1, it means process number 1 has the copy of page number 1. Similarly in row number 1 and column number 2 value 1 means that process number 2 has the copy of the page number 1. Moreover value 1 in row number 4 and column number 2, represents that copy of the page number 4 is also available on process number 2. The values -1 means that the process does not have the copy-set of corresponding pages.

Because we are using just four processes, therefore, we require just four columns to cover all the processes additionally the 5th column is used for the manager of the system if the manager has the copy-set of any page then its value is 1 and the value 0 means that the manager has not the copy-set of the particular page.

```
int [-1,7] iTable[8][7]={  
0,1,1,-1,-1,-1,0,  
0,1,1,-1,-1,-1,0,  
1,1,1,-1,-1,-1,0,  
1,1,1,-1,-1,-1,0,  
2,1,1,-1,-1,-1,0,  
2,1,1,-1,-1,-1,0,  
3,1,1,-1,-1,-1,0,  
3,1,1,-1,-1,-1,0 }
```

This array is page table information maintain on the by the manager. Each row in the array is representing the page number i.e. row number 1 is showing the information about the page number 1 and row number 2 is showing the information about page number 2 and so on for all the other rows in the array. The columns of the array are showing the attributes of the pages which are explained as under.

First column is representing the owner of the page i.e. in row number 1 we have written the 0 in first column which means that the owner of the first page is 0 process and in 7th row there is 3 in first column which shows that the owner of 7th page is process number 3. It is hard coded because in the protocol the ownership is not changed it is only shifted temporarily in write fault case. Second column in the table is representing the information about the access of the page it can be either read or write if the access of the page is marked as read it is made 1 and for write it is made 2 and if the page has both read and write access it is then made 3. Third column in the table is representing the lock information of the page lock. If it is 0 then it means that page is locked and if it is 1 it shows that page is not locked it is unlock and available for operation. Column number 4 to the column number 7 is representing the information about the process having the copy-set of the page i.e. 4th is for first process and 5th is for second if it is 1 against any process this shows that the specific process has the copy-set for that page and if it is 0 it shows that the process has not copy-set of that page. Column number 8 is representing information about either manager has copy-set of the page or not. If it is 1 it shows the manager has the copy-set of that page and if it is 0 it shows that manager has not the copy-set of the process.

```
int [0,7] owner[4][2]={ 0, 0,  
1, 1,
```

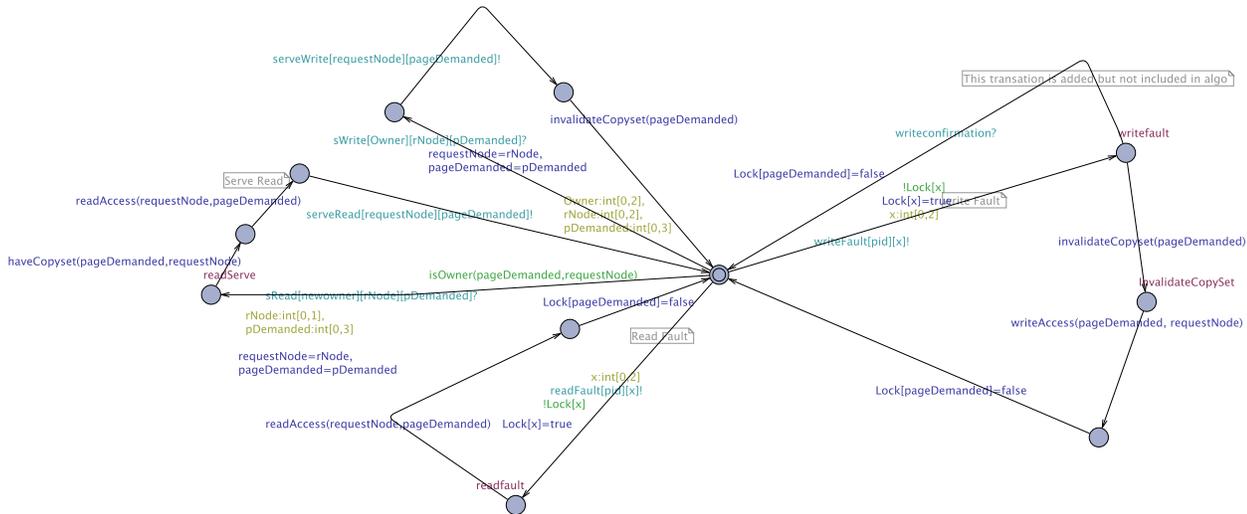


Fig. 6. Improved Central Manager.

```
2, 2,
3, 3 }
```

This array is used to store the information about the ownerships of the pages. In the array row number is representing the number of the page and the column values are representing the owner values of the pages i.e owner of first two pages (page number 1 and page number 2) is process number 0 there for in row number 1 and row number 2 there is 0 values and the owner of page number 3 and 4 is process number 1 there for the values of row number 3 and 4 are 1 and same procedure for remaining pages.

E. Methods or Functions

Following are the methods used:

```
//pid is process id
//x is page number
bool isOwner(int pid,int x)
if((pid == 0 or pid == 1)and x == 0 )
return true;
if((pid == 2 or pid == 3)and x == 1)
return true;
if((pid == 4 or pid == 5)and x == 2 )
return true;
if((pid == 6 or pid == 7)and x == 3 )
return true;
return false;
```

This function is used to find weather the process is owner of the page or not. It is taking process number and page number as arguments and using if statement to compare the value of process with the owner of the page value and if it matches it return true and if it does not match with owner value it returns false as its return type is bool (a Boolean data type). At the last statement if any if statement is not satisfied then it returns false a default value in case of process is not owner of the page. In fact it is using exhaustive search algorithm to find the owner value it is comparing the process and page id with all the possible combinations and then returning the value.

```
void readAccess(int pid, int x)
pTable[pid][x+3] = 1;
```

This function is used to grant access for reading to the requesting process against the required page. It is taking process id and page number as arguments and it is making the change in pTable and changing its access bit discussed earlier in this paper to 1 which means the process have read access of that page. As its return type is void there for it is returning nothing just making the change in pTable.

```
void haveCpoyset( int p, int n)
pTable[p][n+3]=1;
```

This function is used to make the entry of a process in the copy-set of the page. It is taking process number and page number as arguments. This function is returning nothing it is just making the entry in pTable. This function is used in read

operation according to the protocol discussed in this paper if any process has read the page it must have its copy and page table is maintained for this purposes any process which has read the page is entered in the copy-set of the page in pTable.

```
int askForOwner(int p)
if(p == 0 or p == 1)
return 0;
else if(p == 2 or p == 3)
return 1;
return -1;
```

This function is used to find out the owner of the page. It takes page as an argument and returns the owner of the page and returns an integer value which is owner of page which it received. This function is used in read fault and in as well as in read fault. if it receive the page number who has no owner it returns -1. The following method is used to invalidate all copy-set of the page. It takes the page number as an argument and returns nothing. It is making the change in the pTable and changing all the values of the page against copy-set bit to -1. According to the protocol if any process writes the page then it's all copy-sets must be invalidated by the owner of the page and as well as by the manager of the system as an information. Loop is used because in pTable according to the page there is series of bits showing the processes having the copy-sets of that page. We are just putting -1 in place of that process which shows that the page has no copy-set now.

```
void invalidateCopsset(int p)
for(loop=0;loop<4;loop++)
pTable[p][loop+3]=-1;
```

The following method is used in case when the process on which read or write fault occurs is the manager of the system. It is taking the page number as an argument and making the change in copy-set and also in ITable. This method is used only on the manager end and only the information of the page is updated and the values against the manager bits are updated in the tables.

```
void managerUpdat(int p)
copsset[p][4]=1;
iTable[p][6]=1;
```

F. Improved Central Manager Algorithm

The primary difference in central manager and improved central manager is that ownership of page is moved on individual owners so confirmation operation is eliminated.

G. Automaton of Client Process

The automaton of client process is shown in Fig. 2. The client process has 11 states and they are named as initial and it is denoted by double circle and other stages are *read-Fault*, *managerReadFault*, *serveRead*, *confirmation*, *pageLocking*, *readServe*, *accessRead*, *writeFault*, *writeConfirmation* and *requestToOwner*. The initial stage is used to send the request for read fault and write fault request to owner of the page and serve read requests are also generated from initial stage. The major actions performed by the process are described as. The automaton for client process is depicted in Fig. 2. The initial state is named as Client. The client process has two states. The first state is initial state and the second state is committed

state, which is used with action of sending repaired packets to end receivers. There are overall four major actions described as:

- 1) Sending request to read server for read fault of the process against a specific page.
- 2) Receiving request from server as an owner of the page to serve a page for reading.
- 3) Send the page to the requesting node for reading.
- 4) Receive the page form owner for reading purposes.
- 5) Sending the confirmation after receiving the page from owner.
- 6) Receiving manager read fault request is fault is occurred on the manager of the system.
- 7) Sending request to write server for write fault of a process for a page.
- 8) Receiving request from server as an owner of the page to grant access of writing in a page.
- 9) Sending a page for writing purposes to the requesting process.
- 10) Receiving page from owner of the page for reading.
- 11) Sending confirmation to the manager after receiving the page.

The process is using two functions which are described as:

- 1) Changing the access of the page to read for specific process.
- 2) Asking a process to weather it is owner of the page or not.

To communicate with the read server readFault[] channel is used. To communicate with write server writeFault[] channel is used. Client process uses the channel sRead[] receive the request as an owner of the page to serve the page to the requesting process. If the page fault occurs on manager process the managerReadFault[] channel is used by the client process. Client process uses the channel readserve[] to send the required page to the requesting process. Client uses confirmation[] channel to send the read operation completion confirmation to the manager process. Client process uses writeFault[] channel to communicate with write server. Client process uses the channel sWrite [] to receive the request as an owner to serve the page to the requesting process. Client process uses the channel servewrite[] to send the required page to the requesting process. writeConfirmation[] Channel is used by the client process to send the confirmation of completion of write operation. To perform the first step read fault process, the variable !Lock[x] is used to check weather page is locked or not and operation is performed only against unlocked pages. The variable Lock[x]=true is used to lock the page by the client process for which read fault or write fault has been occurred so that it should not be accessed by other process for operations. The variable Lock[pageDemanded]=false is used to unlock the page which is locked before page is served for reading or writing purposes. The function readAccess(requestNode,pageDemanded) is used by the client process (Owner of the page) to grant access for read to requesting process. The variable Node:int[0,2] is used at receiving side and passed to request Node variable and after word it is used to send confirmation message to the manager and also used as a parameter to readAccess(requestNode,pageDemanded) function. The variables rN-

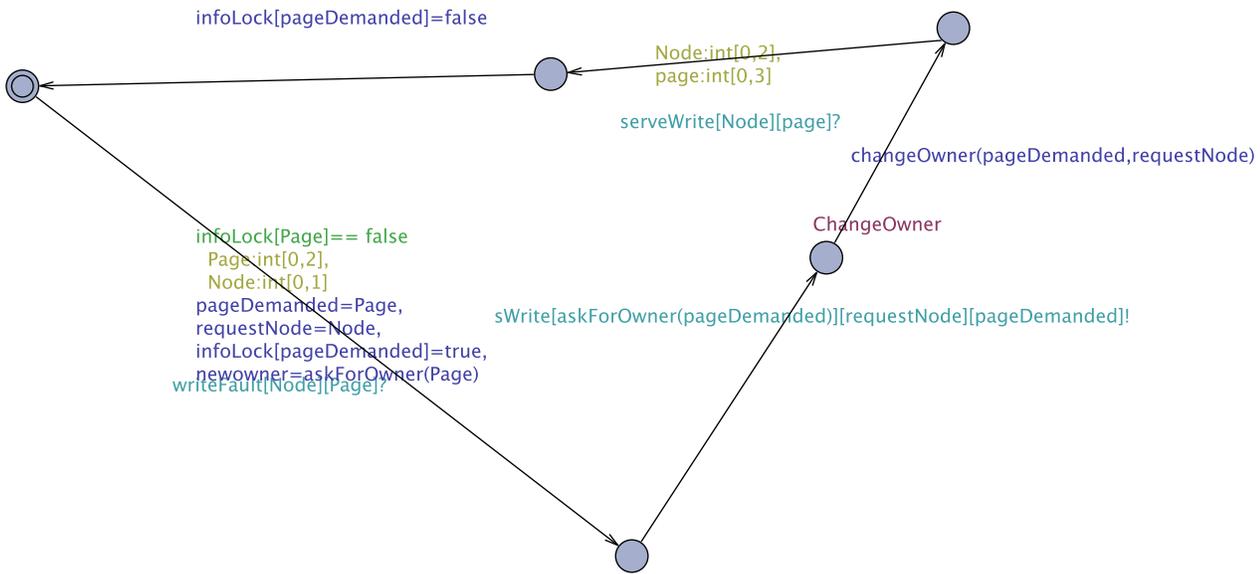


Fig. 7. Improved Write Manager.

ode:int[0,2],pDemanded:int[0,3], Owner:int[0,2] are used on by the owner on receiving side and are passed to request node and page demanded respectively and all of the three variables are used as parameter to sWrite[Owner][rNode][pDemanded]? Channel.

These variables are also used in channel to send page to the requesting process using channel sWrite[Owner][rNode][pDemanded]? For write operation.

In the model shown above the client process is shown in figure. In the first step of read fault is the guard! Lock[x] is used to check the page and insure that it must be unlocked and process locks the page using update Lock[x]=true. In second step of synchronous local variables Owner:int[0,2], rNode:int[0,2], pDemanded:int[0,3] are used and passed to function isOwner(pageDemanded,requestNode) to check the ownership of the page. These variables are also used in the readAccess(requestNode,pageDemanded) function by the owner of the page to grant the access of reading to the requesting process in desired page. In third step of read fault the channel serveRead[requestNode][pageDemanded]!, is used by client to send the required page to the requesting process. in

the fourth step of the client process there are two possible transitions first if requesting process is not manager then the channel serveRead[Node][page]? Is executed and if the requesting process is manager then managerRFault[pid][Node][page]? Channel is executed. The fifth step read fault process in the client process is confirmation[requestNode]! In which requesting process sends the confirmation of receiving the page to the manager of the system and it unlocks the page which was locked at the starting of the operation using the update Lock[pageDemanded]=false.

While in write fault the first step of synchronization process is the guard !Lock[x] that is used to check the page and insure that it must be unlocked and process locks the page using update Lock[x]=true. The local variable x:int[0,3] is used to indicate the process on which write fault can occurred and transaction is performed using writeFault[pid][x]! channel. In second step of the client process synchronization the local variables Owner:int[0,2], rNode:int[0,2] and pDemanded:int[0,3] are used and passed to the sWrite[Owner][rNode][pDemanded]? channel to receive the request from write handler by the owner of page. In third step of the synchronization process of write fault the channel serveWrite[requestNode][pageDemanded]! is

used to send the page to requesting process for writing. At fourth step of write fault process there are two possible options available first if the requesting process is the manager of system then page is received and secondly if requesting node is not the manager then the channel `serveWrite[Node][page]?` is used to receive the page from owner of the page. At fifth step of write fault in client process confirmation message is sent to the manager by requesting node using the channel `confirmation[requestNode]!`, and required page is unlocked using `update Lock[pageDemanded]=false`.

H. Automaton of Read Server

The automaton of read server is depicted in Fig. 3. The initial stage is denoted by double circle. The read server has 8 states. First one initial state and named as `ReadServer` and other states are `handleReadFault`, `InfoLock`, `nonManagerRF`, `RecieveConfirmation`, `InfoLock.padetManager` and `UnLock-Info`. There are over all two functions described as:

- 1) `haveCopysset(pageDemanded,requestNode)`
- 2) `managerUpdat(pageDemanded)`

The read server plays main role between client process and server process. It communicate with client process and server as well. It receives request from client of faulting page and checks the availability of the page. If page is available for the operation then the variables `page:int[0,3]` and `Node:int[0,1]` are initialized and passed to the variables `pageDemanded` and `requestNode` respectively. In the mean while the demanded page is locked using the Boolean variable `infoLock[page]=true`. After this the read server will check the requesting node for manager and if it is not manager of system then it will add the requesting process in the copy-set of the page using the function `haveCopysset(pageDemanded,requestNode)`. After making the transation the information of the page is locked using the Boolean variable `infoLock[pageDemanded]=false`. The channel `sRead[askForOwner(pageDemanded)][request Node]` is used by the read server to send the request the owner of the page asking it to send the required page to the requesting process.

After sending request to the owner of the page the read server will wait for confirmation from the requesting process after receiving the page by using the Chanel `confirmation [requestNode]?` At the receiving end and after receiving the confirmation from the requesting process the information of the page is unlocked using the Boolean variable `infoLock[pageDemanded]=false`. If the requesting node is the manager of the system then at first step the information of the page is locked using the variable `infoLock[pageDemanded]=true`. At the second step the function `managerUpdat(pageDemanded)` is called by the read server to make the entry in the Ptable in the row of the page (as discussed earlier) for the manager having the copy-set of the page. At third step the read server will send the request to the owner of the page to send the requested page to the requesting process (the manager) and unlock the information of the page using the Boolean variable `infoLock[pageDemanded]=false`.

I. Automaton of Write Server

The automaton of write server is shown in Fig. 5. The initial stage is double circled and named as `Start`. The write server

has 8 states. The first one is initial stage and other are `Write`, `copySetInvalidate`, `ManagerConfirmation`, `copySetInvalidate1`, `nonManagerWF`, `RNconfirmation`, `UnLockPage`.

In Fig. 6, the client process of the improved central manager is given. We can see that the main difference between the central manager and the improved central manager is that the later does not send or receive any kind of acknowledgment.

In Fig. 4, the behaviour of read server for the improved central manager is given. It receives a request for page demanding process and operates like a central manager but with the only difference, that is the requesting process does not send or receive any acknowledgment.

In Fig. 7, the model of write server is given. It receives the request from requesting process for writing a page and operates like the central manager's write server but with the only difference, that is the requesting process does not send or receive any confirmations.

IV. CONCLUSION

Memory coherence is a vital issue in today operating system. This paper models the read and write process by using UPPAAL tool. The modeling results highlight the missing details of the read and write protocols. Modeling encounters two types of limitations. First is to limit the number of demand pages, second one to reduce the number of processes. These limitations prevent the system in generating a very huge state space and also in avoiding the state space explosion problem. These limitations are imposed due to limited memory of the machine. The machine can go out of memory during verification phase. Models properties are not affected due to these limitations because a few number of pages and processes can reflect the behavior of a huge system. This small system is transparent reflection of a huge system with maximum number of pages and huge number of processes .

REFERENCES

- [1] C. Baier and J.-P. Katoen, *Principles of Model Checking (Representation and Mind Series)*. The MIT Press, 2008.
- [2] Y. Feng, L. Zhang, D. N. Jansen, N. Zhan, and B. Xia, "Finding polynomial loop invariants for probabilistic programs," in *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*, 2017, pp. 400–416. [Online]. Available: https://doi.org/10.1007/978-3-319-68167-2_26
- [3] K. Li and P. Hudak, "Memory coherence in shared virtual memory systems," in *Proceedings of the Fifth Annual ACM Symposium on Principles of Distributed Computing*, ser. PODC '86. New York, NY, USA: ACM, 1986, pp. 229–239. [Online]. Available: <http://doi.acm.org/10.1145/10590.10610>
- [4] A. David, K. G. Larsen, A. Legay, M. Mikušionis, and D. B. O. Poulsen, "Uppaal smc tutorial," *Int. J. Softw. Tools Technol. Transf.*, vol. 17, no. 4, pp. 397–415, Aug. 2015. [Online]. Available: <http://dx.doi.org/10.1007/s10009-014-0361-y>
- [5] J. Li, L. Zhang, S. Zhu, G. Pu, M. Y. Vardi, and J. He, "An explicit transition system construction approach to LTL satisfiability checking," *Formal Asp. Comput.*, vol. 30, no. 2, pp. 193–217, 2018. [Online]. Available: <https://doi.org/10.1007/s00165-017-0442-2>
- [6] W. Shen, G. Li, C. Lin, and H. Liang, "Foundation of a framework to support compliance checking in construction industry," in *Structured Object-Oriented Formal Language and Method - 7th International Workshop, SOFL+MSVL 2017, Xi'an, China, November 16, 2017, Revised Selected Papers*, 2017, pp. 111–122. [Online]. Available: https://doi.org/10.1007/978-3-319-90104-6_7

- [7] A. Legay, D. Nowotka, D. B. Poulsen, and L. Traonouez, "Statistical model checking of LLVM code," in *Formal Methods - 22nd International Symposium, FM 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 15-17, 2018, Proceedings*, 2018, pp. 542–549. [Online]. Available: https://doi.org/10.1007/978-3-319-95582-7_32
- [8] A. Boudjadar, A. David, J. H. Kim, K. G. Larsen, M. Mikucionis, U. Nyman, and A. Skou, "Statistical and exact schedulability analysis of hierarchical scheduling systems," *Sci. Comput. Program.*, vol. 127, pp. 103–130, 2016. [Online]. Available: <https://doi.org/10.1016/j.scico.2016.05.008>
- [9] A. Ulrich and A. Votintseva, "Experience report: Formal verification and testing in the development of embedded software," in *Proceedings of the IEEE 26th International Symposium on Software Reliability Engineering*, ser. ISSRE '2015. IEEE, 2015, pp. 293–302.
- [10] L. Shan, Y. Wang, N. Fu, X. Zhou, L. Zhao, L. Wan, L. Qiao, and J. Chen, "Formal verification of lunar rover control software using uppaal," in *Lecture Notes in Computer Science*, vol. 8442. Springer, Cham, 2014, pp. 718–732.
- [11] R. Marinescu, H. Kaijser, M. Mikucionis, C. Secleanu, H. Lönn, and A. David, "Analyzing industrial architectural models by simulation and model-checking," in *Formal Techniques for Safety-Critical Systems - Third International Workshop, FTSCS 2014, Luxembourg, November 6-7, 2014. Revised Selected Papers*, 2014, pp. 189–205. [Online]. Available: https://doi.org/10.1007/978-3-319-17581-2_13
- [12] M. Hammad and J. Cook, "Compositional verification of sensor software using uppaal," in *Proceedings of the IEEE 23rd International Symposium on Software Reliability Engineering*, ser. ISSRE '2012. IEEE, 2012.

Digital Technology Disorder: Justification and a Proposed Model of Treatment

Dr Andrew Kear, Sasha L. Folkes

Faculty of Media and Communication, Bournemouth University, UK

Abstract—Due to advances in technology being made at an exponential rate, organisations are attempting to compete with one another by utilising state-of-the-art technology to provide innovative products and services that encourage use. However, there is no moral code to inform sensitive technology design, a consequence of which is the emergence of so-called technology addiction. While addiction as a term is problematic, increasing evidence suggests that related-conditions present implications for the individual, for organisations and for wider society. In this research, a consideration of the potentially addictive elements of technology indicates that it can be possible to reverse engineer these systems, as it were, to promote the development of new behaviours, which can enable the individual to abstain from overuse. Utilising smartphones to deliver digital behavioural change interventions can leverage abundant data touchpoints to provide highly tailored treatment, in addition to allowing for enhanced monitoring and accuracy. To inform understanding of this contemporary phenomenon, the literature on addiction has been reviewed, along with the literature on persuasion architecture to inform an understanding of techniques that lend themselves to overuse and how these can be leveraged to promote recovery. From which, the authors have developed a proposed model to inform the practice of those operating in the domains of computer science.

Keywords—Addiction; digital; treatment; data; smartphone; behaviour; overuse; interventions

I. INTRODUCTION

In the 21st century, there has been a trend with organisations moving away from traditional channels at a rapid rate, opting to invest in their digital counterparts instead. One of the aims of these online channels is to encourage people to spend more time on them, which has been widely successful, resulting in widespread adoption of digital technology devices. Since these systems are growing in importance for both personal and organisational communications, the internet has become one of the most important marketplaces for transactions of goods and services [46], with over “4 billion internet users recorded in 2018” [36]. With usage figures increasing at a rate of “7 per cent year-on-year” [36], there are increased opportunities for consumerism.

Resultantly, “the digital era transcends demographics”: enabling marketers to “have a clearer picture of a consumer from the content they’ve viewed” [88]. Developers are creating highly advanced systems that are “tailored to its host’s needs and reinforcement schedule”, a Skinner box, as it were, that enables individuals to “interact with each other without sacrificing the integrity of their own construct” [21]. Since consumer needs are being met at an accelerating rate, spending

time online can be a highly rewarding experience, “contributing to personal enjoyment for many people” [41].

However, an increasing issue is the overuse of these systems. As noted by [16], “computer systems cannot improve organizational performance if they aren’t used”, however, “complex social situations arise at the individual, organizational, and societal levels” as a result of overuse [14]. Though, who is to blame for overuse? Is it irrational, if immersion, specifically, engagement, is listed as being “a desirable—even essential—human response to computer-mediated activities?” [45]. There is thus a dichotomy between encouraged adoption of technology, and the implications that stem from overuse, with systems listed as “both a creator of certain “dark” effects and a harbinger of their antidotes” [14].

In those that exhibit “problematic levels of usage, devaluation of life itself in the real world is likely to have a major impact on the good life... requiring less effort and providing faster rewards” [41]. Influential systems can encourage “users to spend their time in the virtual realm and abandon reality”, and to “value success and rewards” in alternate realities “as equal as or even higher than those in the real world, devaluing the latter” [41]. As a result of people spending excessive time online within these platforms, negative outcomes such as depressive and anxiety disorders, insomnia and so-called addiction have all been implicated with overuse [28].

These implications can encroach onto an organisational level, associated with “perceived work overload, technology–family conflict”, compromised “organizational commitment” [79] and technostress, which accounts for “50–75 per cent of all information security breaches” in the workplace [14]. While effects such as these are becoming increasingly apparent, the only technology-related condition to be recognised in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), regarded as “the global standard for the classification of mental disorders” is Internet Gaming Disorder (IGD) [24].

Previous attempts at generating official recognition for other technology-related conditions within the DSM-5 have been unsuccessful. Resultantly, there is a lack of funding, research and resources allocated towards identifying new methods of intervention. Currently, there are treatment interventions available for several technology-related conditions. However, the literature suggests that existing treatments can be ineffective. For so-called Internet Addiction and Mobile Phone Addiction, there are detox centres and

military-style camps, which practice abstinence from technology and seek to teach alternate behaviours.

Though effective in select cases, abundant research suggests that these types of treatment can cause more harm than good. In China, “unlicensed training camps” that seek to “wean” individuals off the internet are particularly controversial, having led to cases of casualty and even death [84]. In 2009, “15-year-old Deng Senshan” was found to be dead “just hours after he checked into an Internet bootcamp”, while Pu Liang was admitted “to hospital with water in the lungs and kidney failure” just three weeks after admission [77]. Due to these instances, treatment centres came under scrutiny in 2009, and restrictions were placed on methods used, with laws set.

In 2016, however, reports surfaced of a “16-year-old” starving her mother “to death in revenge for sending her to abusive internet addiction boot camp”, in which “she was beaten and abused” [58]. Since being published, the report led to several “former students” drawing attention to additional cases of “abductions, beatings, and corporal punishment” [58]. Resultantly, consent has become increasingly important in treatment interventions, with collaborative approaches becoming the norm. For example, chemical treatments can enable treatment to be delivered without the individual needing to make drastic lifestyle changes.

Typically, research indicates that anti-depressant drugs allow for both the symptoms of technology-related conditions to subside, along with co-occurring conditions to be targeted [12]. Albeit, “there has been no pharmacological agent identified to be effective” as a stand-alone treatment of technology-related conditions: “all the studies providing an indication for the application of pharmacotherapy... treatment focus first on comorbid disorders” [12]. If technology-related conditions exist in their own right, this would indicate a need for specific treatments to be devised to perhaps ensure a higher rate of recovery.

Further, while contemporary treatments are largely unimposed, there are still several barriers to the effectiveness of treatment. Universal prescription changes can act as a barrier to receiving pharmacotherapy treatments for some, due to personal income levels [37]. Costs are not confined to chemical interventions, however these being “significant barriers to providing evidence-based, behavior change interventions” [49] for healthcare professionals.

For the providers of therapeutic interventions, “financial and staffing resources” for employee training may not always be accessible “to the average community-based treatment program areas” [49]. Additionally, “the squeeze on local authorities, public health budgets” means that “teams are not getting the funding they need”: resultantly, there are extended waiting lists for public treatment, causing “increasing numbers of people to self-refer for costly treatment” [72].

In addition, there are several other elements compromising the effectiveness of treatment: including geographic, political, societal and individual factors. Foremost, distance of patients to treatment centres “may be a barrier to accessing evidence-based care for many patients, especially in rural areas” [49].

Secondly, the time required for treatment interventions can be a deterrent to those with work or family commitments, since “most inpatient facilities insist patients focus completely on treatment” [37]. Finally, there is a stigma attached to addictive disorders, which can deter people from seeking treatment: “almost one-fifth of people who don’t seek treatment say they fear what others would think if they went to rehab” [37].

Due to the issues listed above, it seems appropriate to make a case for new forms of treatment for so-called addictions to be identified. In the case of technology overuse, increasing prevalence rates, widespread adoption and younger device ownership make it seem more important than ever to do so. Before these can be identified, however, the authors wish to make a case for an alternate form of classification to inform treatment. While there are ideas circulating about various technology-related conditions, there is an absence of universally recognised terminology and criteria. In order to appropriately categorise such behaviours, the authors have used the following process:

1) So as not to exclude a category through semantics, ‘digital technology (DT)’ will refer to technology that operates via microprocessors: including computers, applications, the internet, mobile phones, as well as “devices that enable access to cyberspace, the use of digital audio/video and information communications technology (ICT)” [32].

2) In cases of sustained DT overuse that presents above norm negative repercussions, the term ‘Digital Technology Disorder’ (DTD) will be used.

3) While aspects of DTD can share elements of “natural history, phenomenology, tolerance, comorbidity, genetic contribution, neurobiological mechanisms, and response to treatment” with officially recognised addictions [23], ‘disorder’ has been deemed the most appropriate term. This is due to issues associated with ‘addiction’ (see page 3).

4) Since DTD can be compared with IGD, DTD will refer to a “compulsive-impulsive spectrum disorder” [70]. Within which, DTD will refer to cases where DTs are an end in themselves, thus refining the concept further, by excluding cases of DTs being used to enable a pre-existing addiction.

By coining the term, the authors hope to present a case for DTD to be treated as seriously as an officially recognised disorder, which could help towards changing perceptions and identifying new methods of treatment. The authors also wish to point out that, while official recognition of IGD is helpful, it does not sufficiently account for all technology-related conditions. This is due to the fundamental difference: 1) DTs have encroached on almost every aspect of life, with their usage being encouraged in areas such as work and school. Gaming remains a pastime for those who wish to peruse it unless diagnosed with IGD or in cases of employment link explanation.

Due to which, it can be harder to spot those with DTD. There would be no apparent signs, for example, that distinguish a DTD sufferer on the tube to a businessman seeking to meet targets via email. A standardised term would allow people to identify the signs and criteria and to intercept before the repugnant effects are shown. Since there are several research

gaps regarding “how individuals understand, adopt, and learn technology outside of the formal organization” [75], the authors hope to inform sensitive DT design which promotes responsible use. In turn, the authors hope that the future of DT design strives to minimise potentially hazardous outcomes with a consideration of their power.

II. ADDICTION

A. Issues with Addiction

Multiple theories have been proposed to explain addiction. The literature is vast, with key themes of individual responsibility versus individual as a victim [30, 86], rational [9, 51] versus irrational; beneficial versus repugnant, as well as alternate standpoints [22]. Whilst it may seem unproductive to get caught up in semantics, these standpoints are important, since they influence and dictate treatments that are available. As noted by [10], ambiguity “about the meaning of the term “addiction” is not confined to “academic debates... these differing perceptions “matter” in many ways”.

Formerly, perceptions “that are held by scientists and policy makers are likely to influence their policy decisions”, including allocation of resources “for the treatment and prevention of addictive behaviors” [10]. Moreover, attitudes towards addictive behaviours can impinge upon an individual’s understanding of the consequential outcomes and in turn, affect levels of engagement. This may be particularly problematic for vulnerable groups in the population, such as “adolescents, who are at the ages during which many of these behaviors are commonly initiated” [10].

The crucial element here is that prescribed treatment seems to depend, largely, on the perspective. For example, in Asia, the concept of Internet Addiction (IA) is well established, with perceptions of it being a threat to public health. There has been abundant research into prevalence rates, criteria, symptoms, risks and treatments. In 2008 alone, figures estimated that 210,000 children in South Korea (6-18 years old) were undergoing treatment for IA, a number which is predicted to have significantly increased [15]. As “the average South Korean high school student spends about 23 hours each week gaming, another 1.2 million are believed to be at risk for addiction and to require basic counselling” [15].

Consequently, the South Korean government have “trained 1,043 counsellors in the treatment of Internet addiction and enlisted over 190 hospitals and treatment centres”, in addition to setting up “preventative measures” in education [15]. Similarly, in China, it was predicted that “13.7% of Chinese adolescent Internet users meet Internet addiction diagnostic criteria” in 2008, which equates to “about 10 million teenagers” [15]. Thus, the government and legal system in Asia have sought to reduce the interactions with computers among its teenagers, by stigmatising things such as “more than 3 hours of daily game use” [8].

In other parts of the world, such as America, statistics on prevalence rates of similar conditions are lacking, perhaps as a result of differences in public and private consumption. In Asia, internet cafes act as a means of observation for DTD symptoms, however, “in the United States games and virtual sex are accessed from the home” [8]. Further, “attempts to

measure the phenomenon are clouded by shame, denial, and minimization” [8]. As a result, there are fewer treatments available in the US; hence, it seems appropriate to deduce that the different perspectives can be integral to the different treatments prescribed.

Due to differing standpoints, addiction “has been criticized both within and outside the mental health disciplines on a number of grounds” [22]. Addiction is largely subjective: in order to define addiction, “we must select some aspects to include and empathize, and others to exclude” [6]. Prerequisites such as “scientific, guild, societal, and political considerations” both influence and dictate definitions [6], thus addiction is “historically and culturally specific” [6]. Resultantly, contemporary definitions of addiction can be discordant with traditional theories of anthropology, behaviourism and cognitive processing [22, 48].

B. Broadening Scope of Addiction

Prior to the Theory of Rational Addiction [3], “most of the literature in this area until the mid-1980s modelled addiction as habit formation”, specifically, irrational ones [27]. In recent more years, however, the notion that basic activities “can be properly described as addictive” has surfaced [27]. This is because “people get addicted to not only alcohol, cocaine and cigarettes but also to work, eating, music, television, their standard of living, other people” [3], as they conform to “the two conditions required for addiction: reinforcement, and tolerance” [27].

As noted by [30]: “what is coming up fast as being the central core issue ... is continued engagement in self-destructive behaviour despite adverse consequences”. Ideas about the existence of behavioural addictions have surfaced, including “what some are calling positive addictions (exercise, meditation)” [7]. Resultantly, addiction is being attributed to numerous behaviours, resulting in worries that “if what is labelled addiction becomes too broad, the word addiction will become devoid of meaning” [7].

While an increased scope of addictive behaviours could “be justified if common features exist across a similar set of behaviours”, that allow us to better “understand addictive problems and expand society’s capacity to intervene”, there are notable concerns [7]. As illustrated by Heller (2008 cited by) [61]: “if every gratified craving from heroin to designer handbags is a symptom of “addiction,” then the term explains everything and nothing”. Hence, the term “addiction” should be used with caution when referring to behaviours that are not listed as being official within the DSM-5, while a mainstream understanding of the difference between “addictions” and “disorders” would perhaps enable more consistent categorisation.

III. THEORIES OF ADDICTION CREATION

The following section seeks to identify patterns within how addiction is formed. Primarily, theories of addiction creation will be explored, since “the best way to understand and, ultimately, to change addictions is to understand why and how they began” [7].

A. Associative Learning

Though Pavlovian conditioning (Pavlov 1927) and instrumental conditioning (Skinner 1938) are both included within the realm of associative learning, these substrates have crucial differences [52]. Formerly, they differ in terms of experimental grounding. In Pavlov's (1927) original experiment, "the existence of the unconditioned response" was made apparent by "presenting a dog with a bowl of food and the measuring its salivary secretions"; leading to the discovery that "any object or event which the dogs learnt to associate with food (such as the lab assistant) would trigger the same response" [52].

Therefore, Pavlovian conditioning "involves automatic or reflexive responses" to sign-posted stimuli, whereby "conditioned stimuli (CSs) elicit conditioned responses (CRs)" [4]. For example, "anticipatory responses, behavioural habits or even conditioned motivations and emotions" that are "appropriate to the unconditioned reward stimulus (UCS)" [4]. By comparison, instrumental conditioning considers voluntary behaviour. In Skinner's (1938) original experiment, in which rats were observed in The Skinner Box, "actions that are instrumental in gaining access to rewards, such as lever pressing for food" were found to "be controlled by two distinct associative processes" [87].

In the onset, a behaviour is "goal-directed and mediated by the encoding of an association between the action and its specific consequences or outcome", thus, "instrumental performance" can "be sensitive both to non-contingent reward delivery" and to "changes in outcome value" [87]. Over time, however, there is a profound shift. This is due to "control over performance" being "found to shift to a stimulus-response process", whereby "actions become stimulus-bound or habitual", resultantly, actions become de-sensitised "to changes in either the instrumental contingency or reward value" [87].

Types of associative learning also differ in terms of the reward prediction process. Pavlovian conditioning refers to "stimulus-stimulus associations and predictive reward expectation", or "stimulus-response associations and act-outcome representations" [4]. Instrumental conditioning entails "response-contingent reinforcement", whereby "specific instrumental responses are strengthened by response-contingent reinforcement" [4]. Despite these differences, both forms of associative learning are analogous in their application to teaching new behaviours. By understanding both substrates in the context of addiction, it can be possible to deduce several aspects.

Formerly, potential addictions can be conditioned, thus conditioning techniques could be utilised, in turn, to reverse the process and promote new behaviours in the proposed model. For example, having identified in Pavlovian conditioning that sign-posting can elicit an involuntary response, perhaps sign-posting can be used to prevent the formation of potential addictions, by intercepting behaviour at the point of reward expectation. As instrumental conditioning outlines how behaviour is dictated in the onset by associations with actions and the outcome assisted perhaps through educative

intervention at an early stage to create new associations, will provide a greater chance of recovery.

Further, instrumental conditioning bears similarities with O-PT, for its illustration of reward de-sensitisation within addiction formation, and TMRA, with its "diminished euphoric capital" [73]. This could indicate that disorder subjects can be characterised by above normal levels of reward de-sensitisation, hence, this will be examined further in the 'Mind' section. While associative learning has allowed for findings such as these to be extrapolated, there are limitations to their application. For example, instrumental conditioning deals only "with expressible behaviours" [64]. Therefore, there are gaps in understanding inner processes that may play a profound role in addiction development. To uncover these, the next section will explore the mind.

IV. MIND AND NEUROPROCESSING

Despite differences in reward processing, "neural substrates" for both forms of associative learning "are distributed relatively widely across both subcortical and cortical brain structures" [4]. Though psychological components are listed individually, "categories of motivation, learning and emotion or affect constantly interact in reward", with N-S-RDs being "particularly sensitive to different processes" [4]. This indicates that addictions and disorders entail a combination of intertwined processes that require "brain manipulations" to "dissociate many of these processes to reveal their psychological and neural separation" [4].

In particular "manipulations of mesolimbic dopamine systems" have been found to alter "reward 'wanting' without changing reward 'liking'" [5]. Therefore, "neural manipulations could influence rewarded behavior because they alter any one of many forms of learning", the outcome of which will be determined by "precisely which form is altered" [4].

It can be useful to gain an understanding of these systems, to see if it can be possible to utilise this knowledge to promote behavioural shifts within the proposed model. Since this paper seeks to treat an impulsive-compulsive disorder, pathophysiology of officially listed addictions (S-RADs) and officially listed disorders (N-S-RDs) will be drawn upon, so that appropriate conclusions can be made. There are several key similarities to cover, starting with neurocircuitry.

A. Neuroadaptations

A key element of S-RADs is neuroadaptations, which entail a tri-part mental process, including "preoccupation / anticipation, binge intoxication, and withdrawal/negative affect" [43]. This evidence of neuroplasticity has been said to contribute to the transition from a "disorder, that moves from impulsivity to compulsivity", to an addiction [43]. If N-S-RDs can be proven to exhibit similar neuroplasticity, it would indicate a need for time-effective treatment interventions, to intercept behaviour before it is able to develop into an addiction.

Increasing research "suggests that non-drug addictions may lead to neuroadaptations similar to those reported with long-term drug use" [56]. This indicates similar neurocircuitry

regarding “reward processing and decision-making” in both S-RADs and N-S-RDs [86], which could infer that there are certain preconditions that each of these conditions entail. To identify aspects of these, underlying biological mechanisms will be examined.

B. Underlying Biological Mechanisms

In N-S-RDs, similar locations to those for S RADs, for reward processing have been indicated [86]. Thus, urges in S-RADs and N-S-RDs, particularly of the “urge-driven” sort, “may reflect a unitary process” due to similarities in core biological mechanisms [23]. This indicates a need for treatments to integrate aspects of cognition and chemical manipulations to ensure disassociations can be made.

The particular reward pathway (the mesolimbic pathway) [86] indicates that deficiencies in the onset can lead to the development of both conditions, or promote sustained engagement, despite repugnant consequences. If subjects are attempting to self-medicate, this may lend itself towards identifying ways of shifting dependency from maladaptive behaviours to desired behaviours. If promoted behaviours are able to provide rewards that are as effective as existing ones, thus accounting for pre-existing deficits, it could allow an individual to abstain from maladaptive ones. To identify ways of achieving this, specific components of rewards, including implicated neurotransmitters will be examined further.

C. Neurotransmitters

The serotonergic and dopaminergic neurotransmitters associated with S-RADs [23] have been found to be implicated in N-S-RDs [86], with the regulation of “emotions, motivation, decision making, behavioral control” [86] and “inhibition of behavior” [23]; an impairment of serotonin has frequently been linked to S-RADs. Regarding N-S-RDs, impulsivity has been linked to “dysregulated serotonin functioning”, due to “hormonal response after administration of serotonergic drugs” [23].

This indicates that chemical treatments of the serotonergic sort may allow for regulated control within disorder subjects, and therefore, act as a means of achieving behavioural alterations due to enhanced decision-making. While this is promising as a method of treating those with disorders, chemical interventions are costly and therefore may be a barrier to treatment for some. Albeit, the literature suggests that dopaminergic manipulations are possible without medicinal interventions.

Since dopamine is “involved with learning, motivation, and the salience of stimuli, including rewards” [23]; dopamine deficiency has been hypothesised as one of the underlying causes of S-RADs [23] and N-S-RDs [30]. Therefore, “alterations in dopaminergic pathways” are thought to drive subjects with S-RADs and N-S-RDs to repeatedly seek rewards, “that trigger the release of dopamine and produce feelings of pleasure” [23].

Dopaminergic neurons are activated when rewards are received and are heightened through elements of surprise. When rewards become expected, the dopaminergic neurons become emaciated, as they “respond physically to rewards in a manner compatible with the coding of prediction errors” [83].

The principle of instrumental conditioning can lend itself to interpreting this, whereby “learning is blocked when the stimulus is paired with a fully predicted reward” [83].

Resultantly, “the responses of dopamine neurons to conditioned stimuli” can be said to be “governed differentially by the occurrence of reward prediction errors rather than stimulus-reward associations alone” [83]. In addition [83] suggests that when an error is present at the time of the reward, there is enhanced “behavioural and neuronal learning”, a process both exacerbated and encouraged by dopaminergic neurons. Knowledge of dopaminergic neurons will be carried forward and interpreted within this paper, due to its potential to enhance treatment.

V. CONCLUSIONS

To ensure that this lends itself towards developing a model of treatment, several elements have been identified. Having examined the literature on addiction formation and how habits are formed, it seems that the preconditions of potential disorders can include: dopamine deficits, which motivate subjects to engage in certain behaviours over others; above norm levels of reward de-sensitisation which cause repeated reward seeking; above norm need for variable rewards or stimulation to overcome these.

In addition to the aforementioned preconditions, the tools are as follows: unpredictable rewards can lead to habit-formation, which increases motivation to engage in behaviours that enable dopamine deficits to be fulfilled; prediction errors enable enhanced learning, due to dopamine release; behaviours are mediated by associations in the onset of instrumental conditioning. Therefore, any model that seeks to treat disorders should include aspects of “neural manipulations to separate the processes” that are implicated with disorders [4], which may compromise the subject's abilities to make decisions.

For example, with dopaminergic manipulations or signposting to elicit a conditioned response to new behaviours and to form associations that replace existing reward gratification for maladaptive behaviours. In addition, dopaminergic manipulations should be utilised throughout treatment, to prevent relapse by accounting for potential pre-existing dopamine deficits. Further, models should seek to provide rewards of sufficient variance to motivate subjects to participate in sustained treatment over engaging in maladaptive behaviours. Additionally, prediction errors should be utilised to promote enhanced learning, by overcoming the reward de-sensitisation process that contributes to disorders and to ensure that treatments incur maximum effect.

In addition, treatment interventions must be prompt and accessible, to intercept subjects at key points of habit formation, before disorders develop into potential addiction. Finally, educative intervention should be used to change associations that drive behaviour, with guidance on how to develop new coping mechanisms, thus transforming cases of approach-avoidance conflict into cases of approach-approach. To summarise, any model that seeks to treat disorders should cover aspects of motivation, learning and emotion to guide behavioural alterations and separate entwined processes. Relevant findings could lend themselves towards identifying

ways of ascertaining how the mind is processing things whilst subjects are engaged in technology.

Thus, if we want to alter this within treatment, we might want to employ techniques that highlight which areas of the brain are activated and which chemicals are released, post disorder, to see if treatment has been successful. However, there is still a gap in relation to DTD, and how these techniques may lend themselves to treating that. Despite the conclusions drawn, it is not possible to conclude which aspects of technology lend themselves to irrational habit formation. Therefore, section 2 will seek to explore technology techniques and how these contribute to potential disorders. In turn, this will inform the development of a proposed model.

VI. PERSUASION ARCHITECTURE

This section aims to explore technology techniques and how these may lead to addiction. From which, it can be possible to identify how techniques can be leveraged to promote recovery. Primarily, if we can understand systems, we can attempt to understand how an individual might overuse these. Despite rapid advancements in technology being made, “one of the continuing issues of DTs is identifying factors that cause people to accept and make use of systems developed and implemented by others” [39]. Since [38] posited that the solution to effective user interface (UI) design is a “process of manager-analyst interaction in the explication” of assigned models. Though several academics have sought to identify what constitutes acceptance of electronic UI design.

A. User Interface Acceptance

The author in [71] “proposed that a distinction had to be made between technical and organizational validity”, to gain an understanding as to “why systems that met all technical performance standards still were not universally used or understood” [39]. Davis (1985) devised the technology acceptance model (TAM) two years later [15], depicting “the effect of system characteristics on user acceptance of computer-based information systems” [16]. Within TAM, stages can be broken down into design, cognitive, affective and behavioural, with arrows depicting casual relationships.

TAM consists of “two primary predictors— perceived ease of use (EU) and perceived usefulness (U) and the dependent variable behavioural intention (BI)” [39]; assuming that a “potential user's overall attitude toward using a given system” is “a major determinant of whether or not he actually uses it” [11]. Though highly regarded due to its refined style, TAM is not without criticism: “TAM relationships are not borne out in all studies”, due to a “wide variation in the predicted effects in various studies with different types of users and systems” [39].

Consequently, TAM has been adapted by reference [81] to create TAM2, with added elements such as ‘Social Influence’, which entails “three interrelated social forces impinging on an individual facing the opportunity to adopt or reject a new system”. TAM2 presents several implications. For example, past experience also plays a role in leading to perceived

usefulness, intention to use, and consequential use. Hence, if the systems can alter these experiences in the onset, this may be a way of promoting behavioural shifts in treatment.

Further, since voluntariness is listed, it is important for the proposed model to be consensual. In addition, since subjective norm is listed, classifying overuse as DTD could help in shifting perceptions, by making users aware of the implications that stem from overuse. Finally, TAM2 indicates that adequate returns to satisfy stakeholders can still be achieved in cases of technology design that does not seek to encourage overuse. Instead, there is the potential for new business models to be developed, which focus on the ability of the device to act as a mediator of maladaptive behaviour.

This could be used as a key selling point for brands to enhance their image, from which, relevant returns could fuel the development of strategies that ensure that issues that stem from overuse do not affect future society. While TAM2 has been useful in enabling the above to be extracted, it does not detail elements of individual disengagement. As noted by reference [55]: “one area of future research is examining... the specific notion of disengagement... when, if ever, does intense engagement cross the line into addiction?” Hence, the role of engagement will be examined in more detail in the next section, by identifying how much technology leads to behavioural patterns that are irrational.

B. Role of Engagement

Several scholars have listed engagement as a requisite of usability [45]. Listed as the “physical, cognitive, and affective components of user experiences”; engagement entails “media presentation, perceived user control, choice, challenge, feedback, and variety” [55]. In their Proposed Model of Engagement, [55] suggest that engagement and disengagement can occur within the same interaction period, inferring that “engagement itself operates on a continuum...it may be poor, average, or high”.

We can assume, therefore, that to be inflicted with DTD, one must be ranked high and frequent on the engagement scale. Though, high engagement over short periods of time, perhaps when searching for a particular thing online, may not entail DTD: it must be those that score low in disengagement and high on engagement. The Proposed Model of Engagement [55] can also lend itself towards identifying the techniques that encourage excessive use and what techniques reduce that. These will be explored in more detail in the next section.

C. Elements that Promote Engagement

Having previously covered motivation, interest and goals, these will not be examined further. Other elements of engagement, however, should lend themselves towards informing the proposed model, which seeks to reduce overuse. To ensure theory can be translated into practice, proposed elements of engagement, taken from The Proposed Model of Engagement [55] have been compared with the specific techniques that are utilised in DTs to achieve these in Table I.

TABLE I. TECHNIQUES USED TO INCREASE ENGAGEMENT

Engagement Attributes	Techniques Used in DTs to Achieve These
Aesthetic/sensory appeal	Sounds and vibrations that accompany notifications have been found to “align an external trigger (the ping) with an internal trigger (a feeling of boredom, uncertainty, insecurity)”, thus are utilised within systems and applications to promote usage [42]. Similarly, rich colours can be integrated into systems to increase aesthetic appeal [40].
Novelty and attention	Elements such as “variable rewards” are provided to the user, that “are only sometimes distributed”, hence, “the user comes to anticipate the slight rush of the fleeting reward” [17]. Since rewards are unreliable, “the twitchy behavior is triggered”, whereby the user feels compelled “to keep checking for messages, likes, and status updates” [17]. Techniques such as “app notifications, autoplay, likes and messages that self-destruct” have been “scientifically proven to compel” users to engage in DTs: by making users feel as if they are “missing something really important”, systems can motivate users “to watch/check in/respond right now” [42].
Control	Systems can create the illusion of control, to attract users that lack perceived control in real life. Reference [80] note that the common aspects that contribute to this perceived lack of control are: “need for approval (75% for addicts vs. 36% for regular users)”, “fear of rejection (64% vs. 36%)”, “loneliness (64% vs. 38%)”, and “anxiety (64% vs. 36%)”. By enabling the user to progress through stages, work towards rewards, or attain social validation, these illusions of control can be capitalised upon to drive engagement.
Interactivity	Unpredictability can encourage interaction with devices, due to a human tendency to avoid boredom. In 2014, an experiment revealed that “people actually preferred to shock themselves than sit alone with their thoughts for 20 minutes”, indicating that “many of us would prefer chaos over predictability in our daily lives” [25]. By “reinforcing this notion”, social media provides a fluctuating “feedback loop that becomes more arresting the more we use it” [25].
Challenge	Gamification refers to injecting “gameplay elements in non-gaming settings” to “enhance user engagement with a product or service” [76]. Integrating elements such as “leaderboards and badges into an existing system” can enable developers to “tap” into “users’ intrinsic motivations”, in doing so, ignite a desire within the user to participate in the challenge [76].
Feedback	Algorithms can be utilised “to determine the moment” the user “might otherwise walk away”, at which point, certain rewards can be delivered, such as “small wins” to keep the person playing” [51].
Postive affect	Systems can promote “short-term dopamine-driven feedback loops” with likes, shares or interest in online profiles [17]. These so-called “Vanity Metrics” [25] capitalise upon the “goal-obsessed culture” that permeates within contemporary society, leading to users feeling obliged to chase further goals that are encouraged within social media and applications [51].

D. Elements that Promote Disengagement

Having identified that those with DTDs score high on engagement and low on disengagement, frequently and for extended periods of time; external factors will not be examined in more detail. Instead of considering aspects such as system malfunctions or physical interruptions, Table II outlines specific ways of utilising disengagement aspects within the

proposed model. In turn, this should lend itself towards identifying if there can be electronic user interfaces (UIs) that help with DTDs.

TABLE II. DISENGAGEMENT ATTRIBUTES TO INCLUDE IN THE MODEL

Disengagement Attributes	Aspects to Include in Model
Perceived time	Making the user aware of the time spent on their device, with frequent usage reports and/or dashboards that highlight cases of above normal usage. Since DTD individuals can be influenced by the behaviours of others, perhaps reports could detail levels of progress achieved by others (ensuring anonymity, however), which they could aspire to reach.
	Instilling time limits on certain aspects of DTs (that have been identified or consented to by the user as means of achieving their goals).
Negative affect	Shifting colours to grayscale [40], or by reducing sounds and vibrations to reduce aesthetic and sensory appeal.
Disruptions	Intercepting behaviours that indicate deviations from goals, for self-reflection. Since research indicates “that successfully rewiring one’s habits hinges in large part on self-reflection”, it is important that these are provided “in the middle of the habit”, to enable users to “stop and confront” their behaviour, and “consider how” they “might change it” [21].
	Utilising notification tiers, decided collaboratively. For example, there could be levels that dictate which notifications are allowed during downtime, to disrupt the user’s expectation to receive continuous stimulation.
Positive affect	Utilising an AI nurse, to offer methods of coping better in the real world, such as mindfulness, thus reducing dependency on device.
	Providing positive rewards that entail offline activities, thus utilising gamification to encourage the user “become actively interested in attaining goals” [76]. Sending “auto-responses during downtime”, to avoid the user worrying about missing anything important and to generate a sense of community support [63].

E. Alternate Methods of Promoting Disengagement

Since the proposed model seeks to disrupt users flow, findings from neuroscience regarding implicit memory retrieval can lend themselves towards identifying how to activate alternate brain regions. In “An Electrophysiological Signature of Unconscious Recognition Memory”, an experiment [82] found that “people can accurately discriminate repeat stimuli from new stimuli without necessarily knowing it”. In the experiment, participants were shown a collection of “kaleidoscopic images” and asked “to devote their full attention to half of the images, but were distracted by a number task while viewing the remaining half” [47].

Shortly after, participants were asked to “distinguish an image they had previously seen from a new, but very similar picture”: surprisingly, the participants “guessed” correctly more often than they “remembered” correctly” [47]. Thus,

there is evidence that “implicit memory... can occur without the awareness of memory retrieval” [81]. If kaleidoscopic images can promote regions in the brain that an individual is “unconsciously” aware of, this may be a means of the proposed model impacting on the normal flow of activity in the minds of DTD subjects. In turn, this could lead to distraction-conflict. For example, if a series of images were shown throughout phases of treatment, repeated exposure could promote brain regions that are atypical to those activated when using DTs.

F. Choice of Platform for Treatment

The above seems promising for enabling UIs to help treat DTDs. Widespread adoption of smartphones indicates that smartphone-based treatments can “transcend geographic boundaries”, by enabling “on-demand access to therapeutic support outside of formal care settings anytime and anywhere” [50]. In addition, smartphone-based treatments could allow for “linkages to services in one’s community” through GPS [50] to be made. Further, smartphones can reduce costs of treatment. While the “initial development” of smartphone-based “programs can be costly, the cost of hosting and maintaining access to them thereafter” is usually confined “to costs associated with bandwidth needs for deployment and limited technical support” [50].

G. Digital Phenotyping

Since “early recognition of sub-syndromal mood and anxiety symptoms” is paramount in lessening “the pernicious impact of chronic psychological distress and loss of function” [62], there is a need for timely monitoring of disorder subjects’ behaviour. Utilising multifarious digital touchpoints on smartphones can allow for ‘behavioural indicators’ to be collected and interpreted promptly [62]. This is due to digital phenotyping (DP), which involves “moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices” [78]. From which, representations of “a person’s real-time psychological state and overall profile based on their interactions with their smartphone” [13] can be generated.

This is because “digital fingerprints” are thought to “reflect the lived experiences of people in their natural environments” [57], thus giving DP the potential to reveal key behavioural insights on a granular level. Utilising “digital trace data... collected from sensors embedded on mobile devices” [62], DP could gather both active and passive patient data. Active refers to elements that require participation from the user, such as online survey responses or recorded speech samples [78]. For example, speech recordings “could be used to detect vocal markers of mood” [57]. By analysing “lexical content”, “prosody, voice quality” and the “overall tone of voice”, these “vocal cues can provide valuable insights into physical and mental states” [62].

Passive data includes aspects that do not require user participation, such as sleep monitoring, or data acquired from geo-spatial trajectories like the Global Positioning System (GPS). In [62] note that passive data “can enable the identification and tracking” of phenomena such as “mood, fatigue, social connectedness” and “physical isolation”. By monitoring, for example, a fluctuation in sleep levels and

dramatic decreases in movement via GPS, DP “could indicate depression symptoms” [13]. Further, anonymous “call logs” can allow those who fall below the average call reciprocity rate to be identified [62], in turn, allow for the identification of those with an avoidance tendency. Passive data “might also be less susceptible to the complexities introduced by potential linguistic and cultural barriers than more traditional surveys” [57].

Albeit, when passive and active forms are combined, DP could enable the proposed model to not only identify maladaptive behaviours but to promote wellbeing. For example, by providing real-time feedback to the user informing them of their usage patterns, or by highlighting “indices associated with preferred mental states and enhanced behavioural outcomes” [57]. By combining passive and active data with existing medical records, DP can allow for continuous monitoring of behaviour, for inferences in patterns to be made and to ensure that treatment is tailored to individual needs. In addition, it could be possible to “develop precise and temporally dynamic disease phenotypes and markers to diagnose and treat psychiatric and other illnesses” [57].

Thus, DP could enable the model to provide feedback and interruptions at the point of over-engagement, as well as allowing for “the early detection of various conditions” [13]. Since DP does not require “specialized research devices” like other therapeutic interventions such as ecological momentary assessment (EMA), “it is substantially more scalable than traditional implementations” that require “personal digital assistants” [57]. Furthermore, techniques used within treatment can remain up-to-date, by being connected to a cloud-based system; while advances in machine learning can be utilised, such as digital assistants, to provide phone coaching.

H. Neuro-Linguistic Programming and Textual Entailment

Since the proposed model seeks to teach new behaviours, existing techniques from the field of education can be employed. Neuro-Linguistic Programming (N-LP) is used to understand individual differences in thought-processing, methods of communication and “how this communication creates different patterns of behaviour” [33]. N-LP posits “that each individual tends to have a preferred sensory representational system” that is used to “internally code” experiences [74].

By allowing subjects to “adopt an observer position and to interpret their circumstances from a new perspective”, N-LP seeks to “replace dysfunctional thoughts” within subjects that experience “psychological hardship” [69]. N-LP operates under several principles, including the notion that “meaning must match the response” [33]. According to which, both teaching and learning must be a fluid process to ensure that the desired outcome is met by the subject’s behaviour [33]. Fluidity entails variances in “the teaching pattern” employed by the instructor up to the point of generating “the desired response”, in addition to the style of learning used by the learner [33].

Since the proposed model uses DP to attain a stream of real-time active and passive data to provide feedback and alterations in treatment, N-LP is consistent with the authors

understanding of teaching new behaviours. Hence, N-LP will be examined further, to identify techniques that lend themselves towards informing the proposed model. Meta-model questions (M-MQs), for example, can be used to uncover “language patterns that are believed to reflect fundamental cognitive processes” [68]. M-MQs are built on the notion that humans recall events through a “process of generalizing, deleting and distorting information”, which presents implications such as a compromised ability to generate unbiased representations of events [34].

Overtly biased representations of events can impact “the person’s future thinking, core values and health and well-being” [34]. By integrating M-MQs into the proposed model, it could be possible to uncover underlying linguistic processes in subjects, resultantly, to tailor treatment by targeting individual mental models. Advances in technology such as Natural Language Processing (NLP) in Artificial Intelligence (AI) can enable theory to be translated into practice within the proposed model. For instance, by utilising Textual Entailment (TE) models. As noted by reference [19], “TE models take a pair of sentences and predict whether the facts in the first necessarily imply the facts in the second”.

If the proposed model featured aspects of TE models, it could be possible to uncover not just responses to M-MQs, but an individual’s overall response to treatment. Further, periods of relapse could be sign-posted, from which, interceptions can be made in real-time. Forms of TE models can be implemented using a virtual agent, an AI nurse, as it were, to guide subjects through treatment. This will be explored in more detail in the next section, which covers models of computational persuasion, theory and application to the proposed model.

I. Computational Persuasion

As defined in [18], “captology describes the shaded area where computing technology and persuasion overlap”. Since DTs have been found to be detrimental to health in the cases of DTD subjects, an understanding of captology is important in informing the proposed model. Integrating aspects of captology can enable “systems to help people make positive changes to their behaviour, particularly in healthcare and healthy lifestyles” [31]. In particular, by featuring an automated persuasion system (APS) within the model, which “is a system that can engage in a dialogue with a user, i.e. a persuadee, in order to persuade that persuadee to do (or not do) some action or to believe (or not believe) something” [31]. From which, it is conceivable that the model could guide the subjects towards enhanced mental health.

Argument-centric approaches within APS are highlighted [9] to promote behavioural change: “an argument-based approach could be highly beneficial, particularly when someone is lacking some key information, and/or entertaining misconceptions about a topic”. From which, “the system may be able to change the user’s mind about belief in some key arguments”, resultantly, “persuade the user to believe and follow up the persuasion goal” [9]. While the explicit intent is an important consideration, the implicit are also key considerations, since “technology will always be used within a context involving users’ own intentions” [44].

Factors such as “perceived social norms”, “social pressure”, “emotional issues” and “agenda” in addition to “perception of an issue”, “opportunities to change behaviour”, “attitude to persuader” and “attitude to information” [31] all affect the effectiveness of persuasion. In [35] note that there is a “lack of a model which can provide a unified framework for different persuasion strategies... specifically, persuasion is not adaptable to the individuals’ personal states in different situations”.

Thus, [35] propose “a computational model called Model for Adaptive Persuasion (MAP) for virtual agents”, which entails “a semi-connected network model which enables an agent to adapt its persuasion strategies through feedback”. Implementations of the MAP in the form of a virtual nurse were found to be successful in changing the persuades “attitudes and behaviors intentionally, interpret individual differences between users and adapt to user’s behavior for effective persuasion” [35]. Similar to which, the model could feature a MAP-based nurse that utilises data obtained consistently throughout treatment via DP to tailor treatment and make personalised recommendations based on user’s habits and attitudes.

VII. CONCLUSIONS

Having considered the above, it seems appropriate to deduce the following. The stimuli in DTs that can be said to contribute to DTDs include: sounds, rich colours and vibrations (sensory appeal); variable rewards, unpredictability, notifications (novelty and attention); stages to progress through, rewards, social validation (perceived control); fluctuating feedback loops (interactivity); gamification, leaderboards, badges (challenge); wins delivered at point of indicated disengagement (feedback); “short-term dopamine-driven feedback loops” [17], “Vanity Metrics” [25] and goals (positive affect).

These will inform the proposed model, as specific aspects to target. The stimuli also indicate the following conditions of DTD subjects: above norm need for sensory stimulation; reward de-sensitisation, low attention span; lack of perceived control in real life; above norm feelings of loneliness and anxiety; above norm need for intellectual stimulation; above norm need for feedback; pre-existing deficits of dopamine; above norm need for social validation and guidance.

VIII. ANALYSIS OF FINDINGS AND PROPOSED MODEL

Methods of how we learn and become addicted have been explored, along with similarities having been found. In addition, models of effective UI design have been identified that comprehensively outline how systems should be designed that capitalise on human user interaction. Based on findings from sections 1 and 2, it seems appropriate to deduce that the sophistication of technological devices capitalises upon the fragilities of how people learn, to the detriment of human society. Given that DTs are becoming “more sophisticated and their application in different contexts and environments expands” [1], an awareness of the power that technology holds over users is increasingly important.

The literature on addiction and persuasion architecture suggests that the problem of DTD can be derived by computer science techniques manipulating the mind. Due to which, an argument can be made that cognitive solutions are superior to their chemical counterparts. If the systems lend themselves to DTD development, utilising these to promote digital behavioural change interventions can ensure that the issue is targeted at the source, rather than being masked. This may mean that cases of relapse are minimised, in addition to intercepting DTD before it is able to progress into an addiction. Resultantly, there could be a greater chance of recovery, along with minimised negative repercussions.

By “enabling users to face up to their unwanted behaviours as they perform them” [21], cognitive solutions, such as DBCIs, can allow users to achieve a relationship with DTs that is compatible with their desired self. Since integrating aspects of sensitive design into systems can ensure that interactions with DTs are purely productive and cause minimal harm [63], DT-based treatments have the considerable potential of “improving outcomes, reducing costs, and improving the patient experience” [53]. Using smartphones to deliver treatment aspects identified in the proposed model can enable several of the barriers to existing treatments to be removed. By doing so, perhaps more would seek help, in turn; reduce negative implications for the individual, for organisations and for society.

Having identified the potential of utilising smartphone-based treatments, the following section seeks to develop a framework to inform the treatment of DTDs. Formerly, findings from Sections 1 and 2 will be refined, to form DTD preconditions and criteria. Following that, methods of overcoming these will be refined, to include in the model. Then, techniques used in existing treatments will be considered, to ensure the treatment can be optimally effective by addressing barriers to recovery.

A. Synthesis of Findings from Sections 1 and 2

Sections 1 and 2 indicate that the preconditions that DTD is characterised as follows: above norm need for approval; above norm need for sensory stimulation; above norm need for social validation; above norm need for guidance and feedback; above norm need for intellectual stimulation; above norm approach-avoidance attitude; above norm feelings of loneliness and anxiety; above norm reward de-sensitisation that indicates pre-existing deficits of dopamine; above norm lack of perceived control in real life.

Albeit, the above is not enough to diagnose an individual with DTD, these aspects must also meet a set of criteria. These are: cases where there is not an employment link explanation; obsessive characteristics of engagement and time spent online, that are significantly above the norm; non-logical usage: including foregoing rational decisions to achieve irrational high, or losing track of time when online; incurring repugnant consequences; characteristics must be exhibited for extended periods of time, with the tendency for time and needs to increase. Methods of overcoming preconditions listed above, using DBCIs are listed in Table III.

TABLE III. ASPECTS TO INCLUDE IN THE MODEL OF TREATMENT

DBCI	Aspects to Include in Model
Percieved time	Feedback, to inform of usage and progress, such as time spent on device and how time compares with norms to sign-post behaviour.
Negative affect	Grayscale during periods of relapse and time limits on usage (collaboratively decided).
Disruptions	Interruptions during deviations from goals to allow for self-reflection. Notification tiers (collaboratively decided), including restricted sounds and vibrations, to filter distractions from achieving goals.
Positive affect	Gamification of goals, to encourage the subject to progress through treatment, addressing need to be stimulated intellectually. Short-term, offline rewards when goals are met, to reduce dependence on online rewards, to fulfill dopamine deficit and to condition the subject to look forward to offline rewards. AI nurse to reduce feelings of loneliness and anxiety, by guiding and supporting, thus reducing long-term reliance on DTs. Auto-responses to contacts (collaboratively decided), to reduce anxiety towards missing something important when attempting to meet goals. Linkages to community can be promoted through digital phenotyping to encourage integration. Promotion of coping mechanisms to achieve offline sense of control (grounded in psychological, scientific and technological theory and best practice), that is constantly updated, from a cloud-system. This ensures that control is more than illusory, and translates into methods of coping better in the real world, while allowing for new associations to be made.

Aspects in Table III must also meet the following criteria: treatment must not be imposed, it must be desired the by individual who wishes to develop new behaviours; aspects of proposed DBCIs and aims of treatment must be detailed clearly prior to and during treatment, and agreed to by the user; data must not be shared without consent, and must be securely processed to protect the individual and any concerns identified in self-report data must be addressed.

B. Existing Treatment Techniques to Enhance Model

Since integrated approaches “tend to be successful in treating addiction and maintaining recovery” [60], the following section will explore techniques utilised within existing treatments. From which, it should be possible to extrapolate best practice to inform the proposed model. As the literature suggests that any model that seeks to change behaviour faces issues of “motivation, ambivalence and resistance” [29], specific techniques used to target these will be explored.

C. Motivation as an Issue

Motivation in the context of psychotherapy usually refers to avoidance motivation that occurs within the subject undergoing therapy. Avoidance motivation stems from a human desire “to evade unpleasant experiences”, namely, experiential avoidance, and is thought “to protect the individual from re-experiencing past adversities” [26]. While helpful in select

cases, an excess of avoidance motivation can contribute “to the development and maintenance of many psychopathological disorders” [26]. Since a contributing factor to the maintenance of avoidance motivation is “mental representations of undesired transactions with the environment”, methods of restructuring these will be examined [26].

D. Methods of Targeting Avoidance Motivation

Having identified that DTD subjects may use approach-avoidance systems, approach-avoidance theories of motivation will be covered first. These “are active in the presence of potential positive or negative outcomes, respectively” [59]. To increase perceived positive outcomes, groups can be used, since they “are associated with rewards... a source of security and thought to reduce the occurrence of negative outcomes” [59].

As the proposed model seeks to shift behaviour from maladaptive use to purely productive usage, “systematic group influences on mood, information processing, perceptions, attention, and behavior” could be utilised, in the form of “group membership and interaction”, which, in turn, “should impact activation of group members’ approach and avoidance motivation systems” [59]. Hence, the community linkages made possible via DP within the model should allow for systematic group influences on the subjects behaviour, by increasing perceived positive outcomes.

E. Ambivalence as an Issue

Ambivalence is when individuals find it increasingly difficult to change, despite their best intentions, due to being “pulled in two directions by motivations to change and motivations to maintain the status quo” [2]. In cases where maladaptive behaviours “serve important functions”, such as providing an escape from stressful day-to-day life [2], ambivalence is common.

F. Methods of Overcoming Ambivalence

A collaborative technique used in psychotherapy to overcome ambivalence is motivational interviewing (MI), which seeks to increase a “patients’ own motivation for change and adherence to treatment” [66] and reduce resistance [29]. In essence, MI is “collaborative, evocative, and honouring of patient autonomy” [66], as it seeks to guide according to the subject’s personal goals. By following “the acronym RULE”, MI seeks to “Resist the righting reflex; Understand the patient’s own motivations; Listen with empathy; and Empower the patient” [29].

MI has been found to be “equivalent to or better than other treatments such as cognitive behavioural therapy (CBT) or pharmacotherapy” when treating S-RADs, in addition to having “been shown to be efficacious” in treating several N-S-RDs, such as IGD [66] and hypersexual disorder [29]. To incur maximum effect, MI requires “someone with the skill to switch among communication styles and the wisdom to seek and understand what style the patient needs” [66].

Having covered TE models on page 9, it is conceivable that the AI nurse could effectively conduct MI in the first phase of treatment. This could enable learner-centred objectives to be attained, from which, the AI nurse could measure the patient’s

progress via DP throughout treatment. Deviations could be actioned upon, with elements of persuasion architecture being used in relation to the subject’s progress. MMQs could also be used within MI, to decipher the subject’s mental models. Self-report data throughout treatment could measure the effectiveness of treatment, by comparing initial response styles to later styles.

In addition to MI, the Principle of Effective Facilitation (PEF) could be utilised to help overcome ambivalence. The Principle of Effective Facilitation (PEF) is derived from humanistic approaches to treatment. In short, PEF entails external validation to encourage the learner to progress through stages of learning. PEF operates under the following assumption: when individuals “decide to learn”, they “invest time and energy in checking the potential benefits” [67]. If the individual believes that they are positively altering their behaviour in a way that is compatible with the desired self, they will, according to humanism, seek to engage in learning.

Since individuals have a propensity to “devalue their work if not validated”, the use of an “external authority” can allow for enhanced learning within PEF. By guiding the individual through various stages of learning, validating their progress, the likelihood of learning new behaviours should be increased. To incur maximum effect, the external authority should “encourage the adult learner to interact with his environment on his own terms” [67]. By identifying learner-centred objectives within MI, this should enable the AI nurse, who acts as the external authority “to facilitate self-directed learning” [67]. For example, with guidance grounded in theory and usage reports.

G. Resistance as an Issue

Resistance is the “overt or covert reluctance to change and grow” in subjects of therapeutic intervention [54]. Judgment or “adverse emotional reactions” in the therapist “unwittingly may exacerbate the client’s resistance”, while “gaining an accurate, empathic understanding of the client’s difficulties in changing” can reduce resistance [54]. Approaches to treatment that seek to avoid resistance should feature both “active involvement in educating and stimulating the client toward greater awareness and adaptability” as well as allowing for degrees of self-directed learning [54].

H. Methods of Overcoming Resistance

To account for the above, psychoeducation could be utilised within the model. Psychoeducation involves “the provision of information, in a variety of media, about the nature of stress, posttraumatic and other symptoms, and what to do about them” [85]. In several cases, psychoeducation has been found to increase adherence to treatment and reduce relapse rates in subjects with S-RADs and N-S-RDs [65]. Having identified that DTD subjects are not always aware of the negative implications of overuse, psychoeducation could be used to fill in conceptual gaps and to provide coping mechanisms.

In the initial phase, following MI, psychoeducation could be used to educate the user of the potential outcomes of treatment and consequences of overuse. In later stages, psychoeducation could be used at times of indicated relapse, to

reinforce the subject’s motivation to participate in sustained treatment. Further, within phases of psychoeducation, mindfulness-based interventions (MBIs) could be utilised. In the past decade, MBIs have grown in popularity, due to their promising application to treating S-RADs and N-S-RDs [20].

MBIs have been found to be “successful for reducing dependence, craving, and other addiction-related symptoms by also improving mood state and emotion dysregulation” [69]. For example, by minimising “substance misuse and craving by modulating cognitive, affective, and psychophysiological processes integral to self-regulation and reward processing” [20]. As a result, it is hypothesised that MBIs have the potential to “reverse the allostatic process by which normal reward learning is usurped by addictive substances” [20].

Though studies largely relate to S-RADs, aspects of MBIs can be integrated into the model of DTD treatment, since “MBT is an effective treatment for a variety of psychological problems, and is especially effective for reducing anxiety, depression, and stress” [37]. Having identified that DTs can produce rewards comparable to substances, utilising MBIs in the treatment could enable the subject to potentially shift dependence on online rewards to offline rewards. This is due to “the restructuring reward hypothesis”, which “states that mindfulness may reduce addictive behavior by shifting the relative salience of drug and natural rewards from valuation of drug-related reward back to valuation of natural rewards that were salient before the development of addiction” [20].

Due to their ability to achieve “attention regulation and positive affect”, it is conceivable that MBIs “might nonetheless increase pleasure from perceptual and sensorimotor experiences in a fashion similar to sensate-focus techniques and promote positive emotion regulation by amplifying selective attentional processes” [20].

IX. PROPOSED MODEL AND CONCLUSIONS

For example, with “Mindfulness-Based Relapse Prevention (MBRP) and Mindfulness-Oriented Recovery Enhancement

(MORE)” that “have been tailored to directly to address the mechanisms that undergird addiction” [20]. Thus, utilising MBIs within psychoeducation could potentially enable the DTD subject to gravitate towards approach-approach as opposed to approach-avoidance.

A. Response Tendency Hierarchy

These elements should transform non-logical usage into logical, productive usage. Further, previous cases of no employment link explanation should become mostly employment link explanation. Characteristics, time and needs should return to the norm, with the overall positive effect leading to less reliance on DTs. Finally, the elements discussed and explored combine to contribute to Fig. 1 (The Tri Path model of treatment for Digital Technology Disorder).

IMPLICATIONS AND AREAS FOR FUTURE RESEARCH

The implications of this paper include health provision, treatment, corporate social responsibility, better health, wellbeing and a reduction in DTD. Thus, the authors recommend the following areas for further research: a large scale quantitative research to find the most influential path in the proposed model. Research on image kaleidoscoping regarding DTD is needed, in particular kaleidoscopic images from relevant domains of interest may provide some important answers pertaining to this vain of treatment. Furthermore age-specific research regarding DT use and disorders will address a significant research gap. Finally, utilising neuroscience to map the different phases of DTD when completing the varied phases of treatment.

Finally, the authors recognise that some people will be able to self-regulate, after/or during exposure to the DTD treatment proposed. Albeit, the authors also recognise that some people with a high propensity of addiction may need long term support from the AI nurse. Hence, further research may aim to explore time scales for different propensities of addiction.

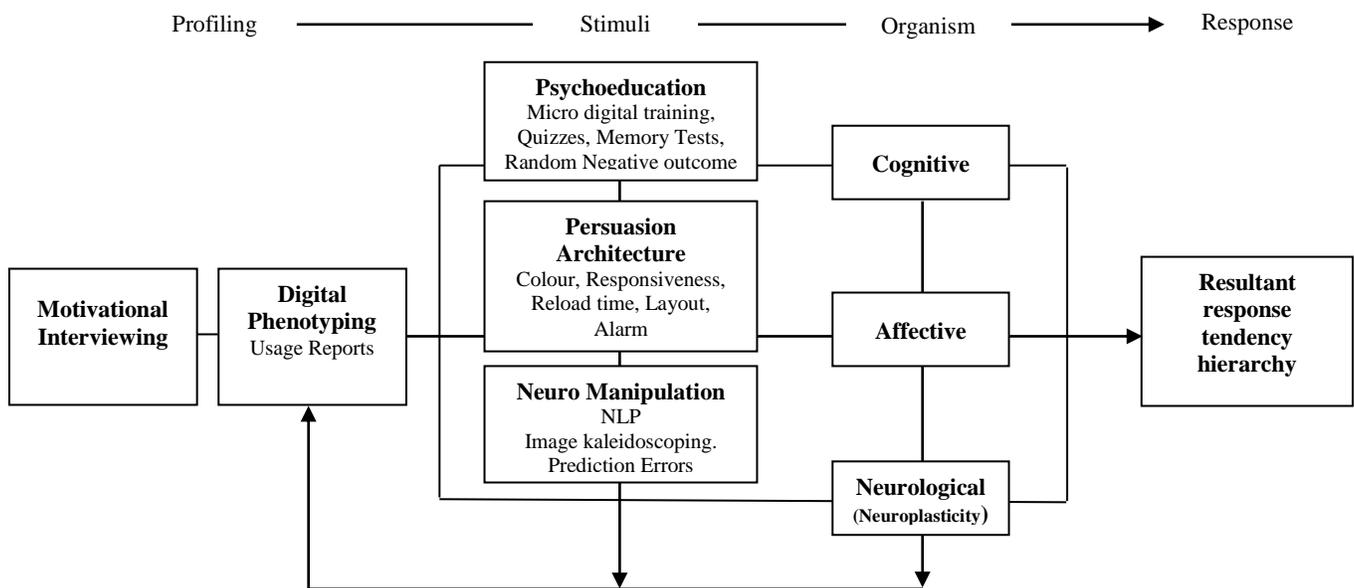


Fig. 1. Tri Path Model of Treatment for Digital Technology Disorder

REFERENCES

- [1] Allen, C., 2006. Why Machine Ethics? *IEEE Intelligent Systems* [online], 21 (4), 12-17.
- [2] Arkowitz, H. and Lilienfeld, S., 2007. Why Don't People Change? *Scientific American Mind* [online], June/July 2017. Available from: http://faculty.fortlewis.edu/burke_b/counseling/cs%20readings/sciam-why%20people%20change.pdf [Accessed 6 July 2018].
- [3] Becker, G. and Murphy, K., 1988. A Theory of Rational Addiction. *Journal of Political Economy* [online], 96 (4), 675-700.
- [4] Berridge, K. and Robinson, T., 2003. Parsing reward. *TRENDS in Neurosciences* [online], 26 (9), 507-513.
- [5] Berridge, K. Robinson, T. and Aldridge, W., 2009. Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion Pharmacology* [online], 9 (1), 65-73.
- [6] Bickel, W. Mueller, T. and Jarmolowicz, D., 2013. What is Addiction? In: Epstein, E. *Addictions: A Comprehensive Guidebook*. USA: OUP. 1-13.
- [7] Blane H. and Kosten T., 2002. Seeking Safety: A Treatment Manual for PTSD and Substance Abuse. New York: The Guilford Press.
- [8] Block, J., 2008. Issues for DSM-V: Internet Addiction. *American Journal Psychiatry* [online]. 165 (3), 305-306.
- [9] Chalaguine, L. Hadoux, E. Hamilton, F. Hayward, A. Hunter, A. Polberg, S. and Potts, H., 2018. Domain Modelling in Computational Persuasion for Behaviour Change in Healthcare. *Computer Science > Artificial Intelligence* [online]. Available from: <https://arxiv.org/abs/1802.10054v1> [Accessed 23 July 2018].
- [10] Chassin, L. Presson, C. Rose, J. Sherman, S., 2007. What is addiction? Age-related differences in the meaning of addiction. *Drug and Alcohol Dependence* [online], 87 (1), 30-38.
- [11] Chuttur, M., 2009. Overview of the Technology Acceptance Model: Origins, Developments and Future Directions. *Sprouts: Working Papers on Information Systems* [online], 9 (37), 1-21.
- [12] Chwaszcz, J., 2016. Gambling and Internet addictions - epidemiology and treatment [online]. In: Chwaszcz, J. and Lelonek-Kuleta, B. *Gambling and Internet addictions - epidemiology and treatment*. France: Natanaelum Association Institute for Psychoprevention and Psychotherapy, 56-58.
- [13] D'Alfonso, S. and Alvarez, M., 2017. *4 WAYS TECH CAN HELP YOUR MENTAL HEALTH* [online]. Melbourne: Pursuit. Available from: <https://pursuit.unimelb.edu.au/articles/4-ways-tech-can-help-your-mental-health> [Accessed 1 May 2018].
- [14] D'Arcy, J. Gupta, A. Tarafdar, M. and Turel, O., 2014. Reflecting on the "Dark Side" of Information Technology Use. *Communications of the Association for Information Systems* [online], 35 (5), 110-118.
- [15] Davis, F., 1986. *A technology acceptance model for empirically testing new end-user information systems : theory and results* [online]. Thesis (PhD). Massachusetts Institute of Technology.
- [16] Davis, F. Bagozzi, R. Warshaw, P., 1989. User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science* [online], 35 (8), 982-1003.
- [17] Dillard-Wright, D., 2018. *Technology Designed for Addiction* [online]. United States: Psychology Today. Available from: <https://www.psychologytoday.com/gb/blog/boundless/201801/technology-designed-addiction> [Accessed 1 May 2018].
- [18] Fogg, BJ., 1998. Persuasive Computers: Perspectives and Research Directions. *Human-Computer Interaction* [online], 225-231.[
- [19] Gardner, M. Grus, J. Neumann, M. Tafjord, O. Dasigi, P. Liu, N. Peters, M. Schmitz, M. and Zettlemoyer, L., 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. *arXiv* [online]. Available from: <https://arxiv.org/pdf/1803.07640.pdf> [Accessed 19 June 2018].
- [20] Garland, E. and Howard, M., 2018. Mindfulness-based treatment of addiction: current state of the field and envisioning the next wave of research. *Addiction Science & Clinical Practice* [online], 13 (14), 1-14.
- [21] Gonzalez, R., 2018. Can our phones save us from our phones? *Wired* [online], 30 January 2018. Available from: <https://www.wired.com/story/can-our-phones-save-us-from-our-phones/> [Accessed 11 March 2018].
- [22] Goodman, A., 1990. Addiction: definition and implications. *British Journal of Addiction* [online], 85 (1), 1403-1408.
- [23] Grant, J. Potenza, M. Weinstein, A. and Gorelick, D., 2011. Introduction to Behavioural Addictions. *The American Journal of Drug and Alcohol Abuse* [online], 36 (5), 233-241.
- [24] Grecu, C., 2013. THE NEW GENERATIONS AND THE ADDICTIONS TO TECHNOLOGY. *European Journal of Science and Theology* [online], 9 (1), 99-110.
- [25] Greer, S., 2018. The Addiction Economy. *Medium* [online], 12 February 2018. Available from: <https://medium.com/@scottcgreer/how-do-we-stop-technology-addiction-c0c081b8c970> [Accessed 1 March 2018].
- [26] Grosse Holtforth, M. Grawe, K. Castonguay, L., 2006. Predicting a reduction of avoidance motivation in psychotherapy: Toward the delineation of differential processes of change operating at different phases of treatment. *Psychotherapy Research* [online], 16 (5), 639-644.
- [27] Gruber, J. and Koszegi, B., 2018. IS ADDICTION "RATIONAL"? THEORY AND EVIDENCE. *The Quarterly Journal of Economics* [online], 116 (1), 1261-1264.
- [28] Ha, J-H. Yoo, H-J. Cho, I-H. Chin, B. Shin, D. and Kim, J-H., 2006. Psychiatric comorbidity assessed in Korean children and adolescents who screen positive for Internet addiction. *Journal of Clinical Psychiatry* [online], 67 (5), 821-826.
- [29] Hall, K. Gibbie, T. and Lubman, G., 2012. Motivational interviewing techniques: Facilitating behaviour change in the general practice setting. *Australian Family Physician* [online], 41 (9), 660-667.
- [30] Holden, C., 2001. 'Behavioral' addictions: Do they exist? *Science* [online], 294 (1), 980-982.
- [31] Hunter, A., 2016. Computational Persuasion with Applications in Behaviour Change. *Frontiers in Artificial Intelligence and Applications* [online]. 287 (1), 5-18.
- [32] IGI Global, 2018. *What is Digital Technology* [online]. Hershey: IGI Global. Available from: <https://www.igiglobal.com/dictionary/digital-technology/7723> [Accessed 1 May 2018].
- [33] Ilyas, M., 2017. Finding Relationships between Acquisition of Basic Skills and Neuro-linguistic Programming Techniques. *Journal of Literature, Languages and Linguistics* [online], 34 (1), 22-25.
- [34] Johnson, MI. and Hudson, M., 2016. Generalizing, deleting and distorting information about the experience and communication of chronic pain. *Pain Management* [online], 6 (5), 411-414.
- [35] Kang, Y. Tan, A-H. and Miao, C., 2015. An Adaptive Computational Model for Personalized Persuasion [online]. In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Argentina July 25 - 31 2015*. Buenos Aires: International Joint Conference on Artificial Intelligence (IJCAI). Available from: <https://dl.acm.org/citation.cfm?id=2832258> [Accessed 23 June 2018].
- [36] Kemp, S., 2018. *DIGITAL IN 2018: WORLD'S INTERNET USERS PASS THE 4 BILLION MARK* [online]. New York: We Are Social. Available from: <https://wearesocial.com/blog/2018/01/global-digital-report-2018> [Accessed 2 May 2018].
- [37] Khoury, B. Lecomte, T. Fortin, G. Masse, M. Therien, P. Bouchard, V. Chapleau, M-A. Paquin, K. Hofmann, S., 2013. Mindfulness-based therapy: A comprehensive meta-analysis. *Clinical Psychology Review* [online], 33 (1), 763-771.
- [38] King, W. and Cleland, D., 1971. Manager-analyst teamwork in MIS: Cooperation vital in systems design. *Business Horizons* [online], 14 (2), 59-68.
- [39] King, W. and He, J., 2006. A meta-analysis of the Technology Acceptance Model. *Information & Management* [online] 43 (6), 740-755.
- [40] Kirsch, M., 2017. *Change Your Screen to Grayscale to Combat Phone Addiction* [online]. United States: Lifehacker. Available from: <https://lifehacker.com/change-your-screen-to-grayscale-to-combat-phone-addicti-1795821843> [Accessed 11 March 2018].
- [41] Klemm, C. and Pieters, W., 2017. Game mechanics and technological mediation: an ethical perspective on the effects of MMORPG's. *Ethics Information Technology* [online], 19 (1), 81-93.
- [42] Knorr, C., 2017. How to resist technology addiction. *CNN* [online], 9 November 2017. Available from:

- <https://edition.cnn.com/2017/11/09/health/science-of-tech-obsession-partner/index.html> [Accessed 11 March 2018].
- [43] Koob, G. and Simon, E., 2009. The Neurobiology of Addiction: Where We Have Been and Where We Are Going. *J Drug Issues* [online], 39 (1), 115-132.
- [44] Krischkowsky, A. Maurer, B. Tscheligi, M., 2016. Captology and Technology Appropriation: Unintended Use as a Source for Designing Persuasive Technologies [online]. In: Meschtscherjakov, A. De Ruyter, B. Fuchsberger, V. Murer, M. and Tscheligi, M. *Persuasive Technology*. Germany: Springer, 78-83.
- [45] Laurel, B., 1993. *Computers as theatre*. 2nd Edition. Reading: Addison-Wesley.
- [46] Leeflang, P. Verhoef, P. Dahlström, P and Freundt, T., 2014. Challenges and solutions for marketing in a digital era. *European Management Journal* [online], 32 (1), 1-12.
- [47] Ledford, H., 2009. Hidden memories guide choices. *Nature* [online], 9 February 2009. Available from: <https://www.nature.com/news/2009/090209/full/news.2009.88.html> [Accessed 19 June 2018].
- [48] Leshner, A., 1999. Science is Revolutionizing Our View of Addiction— and What to Do About It. *The American Journal of Psychiatry* [online], 156 (1), 1-3.
- [49] Marsch, L. and Dallery, J., 2012. Advances in the Psychosocial Treatment of Addiction: The Role of Technology in the Delivery of Evidence-Based Psychosocial Treatment. *Psychiatry Clinical North America* [online], 35 (2), 481–493.
- [50] Marsch, L., 2013. LEVERAGING TECHNOLOGY TO ENHANCE ADDICTION TREATMENT AND RECOVERY. *Journal of Addictive Diseases* [online], 31 (3), 313-318.
- [51] McBain, S., 2017. How technology companies are keeping you addicted to your phone. *New Statesman* [online], 6 March 2017. Available from: <https://www.newstatesman.com/culture/books/2017/03/how-technology-companies-are-keeping-you-addicted-your-phone> [Accessed 1 March 2018].
- [52] McLeod, S., 2013. *Skinner - Operant Conditioning* [online]. United States: Simply Psychology. Available from: <https://www.simplypsychology.org/operant-conditioning.html> [Accessed 1 May 2018].
- [53] Michie, S. Yardley, L. West, R. Patrick, K. and Greaves, F., 2017. Developing and Evaluating Digital Interventions to Promote Behavior Change in Health and Health Care: Recommendations Resulting From an International Workshop. *Journal of Medical Internet Research* [online], 19 (6), 1-11.
- [54] Newman, C., 2002. A Cognitive Perspective on Resistance in Psychotherapy. *Psychotherapy in Practice* [online], 58 (2), 165–174.
- [55] O'Brien, H. and Toms, E., 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science & Technology* [online], 59 (6), 938-955.
- [56] Olsen, C., 2011. Natural Rewards, Neuroplasticity, and Non-Drug Addictions. *Neuropharmacology* [online], 61 (7), 1109–1122.
- [57] Onnela, J-P. and Rauch, S., 2016. Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology* [online], 41 (1), 1691–1696.
- [58] Osborne, S., 2016. *Chinese teen starves mother to death in revenge for sending her to abusive internet addiction boot camp* [online]. London: Independent Digital News & Media. Available from: <https://www.independent.co.uk/news/world/asia/chinese-teen-starves-mother-to-death-revenge-internet-addiction-boot-camp-a7329351.html> [Accessed 20 June 2018].
- [59] Park, E. and Hinsz, V., 2006. “Strength and Safety in Numbers”: A Theoretical Perspective on Group Influences on Approach and Avoidance Motivation. *Motivation and Emotion* [online], 30 (1), 135-142.
- [60] Patterson, E., 2015. *Addiction Treatment Therapies: An Overview* [online]. San Diego: Recovery Brands. Available from: <https://drugabuse.com/library/addiction-treatment-therapies-an-overview/> [Accessed 1 May 2018].
- [61] Pies, R., 2009. Should DSM-V Designate “Internet Addiction” a Mental Disorder? *Psychiatry (Edgmont)* [online], 6 (2), 31-37.
- [62] Place, S. Blanch-Hartigan, D. Azarbayejani, A., 2017. Behavioral Indicators on a Mobile Sensing Platform Predict Clinically Validated Psychiatric Symptoms of Mood and Anxiety Disorders. *Journal of Medical Internet Research* [online], 19 (3), 1-21.
- [63] Price, C., 2018. 9 Ways to Finally Stop Spending So Much Time on Your Phone. *Time* [online], 8 February 2018. Available from: <http://time.com/5139859/smartphone-addiction-solutions/> [Accessed 2 May 2018].
- [64] Psychestudy, 2018. *Classical vs Operant Conditioning* [online]. New York: Psychestudy. Available from: <https://www.psychestudy.com/behavioral/learning-memory/classical-conditioning/classical-vs-operant> [Accessed 11 March 2018].
- [65] Ran, MS. Xiang, MZ. Chan, CLW. Leff, J. Simpson, P. and Huang, MS., 2003. Effectiveness of psychoeducational intervention for rural Chinese families experiencing schizophrenia: a randomised controlled trial. *Social Psychiatry and Psychiatric Epidemiology* [online], 38 (2), 69–75.
- [66] Rollnick, S. Miller, W and Butler, C., 2008. *Motivational interviewing in health care. Helping patients change behavior*. New York: The Guilford Press.
- [67] Rostami, K. and Khadjooi, K., 2010. The implications of Behaviorism and Humanism theories in medical education. *Gastroenterology and Hepatology* [online], 3 (2), 65-70.
- [68] Sancho, M. De Gracia, M. Rodríguez, R. Mallorquí-Bagué, N. Sánchez-González, J. Trujols, J. Sánchez, I. Jiménez-Murcia, S. and Menchón, J., 2018. Mindfulness-Based Interventions for the Treatment of Substance and Behavioral Addictions: A Systematic Review. *Frontiers in Psychiatry* [online], 9 (95), 1-9.
- [69] Savardelavar, M. and Kuan, G., 2017. The use of neuro-linguistic programming as an educational-therapeutic programme: two case studies. *Education in Medicine Journal* [online], 9 (1), 49–58.
- [70] Shapira, N-A. Lessig, M-C. Goldsmith, T-D., 2003. Problematic internet use: proposed classification and diagnostic criteria. *Depression and Anxiety* [online], 17 (1), 207-216.
- [71] Schultz, R.L. and Slevin, D.P., 1983. The implementation profile. *Interfaces* [online], 13(1), 87–92.
- [72] Siddique, H., 2018. UK addiction treatment centres braced for new year surge in demand. *The Guardian* [online], 2 January 2018. Available from: <https://www.theguardian.com/society/2018/jan/02/uk-addiction-treatment-centres-braced-for-new-year-surge-in-demand> [Accessed 1 May 2018].
- [73] Stigler, G. and Becker, G., 1977. De Gustibus non est Disputandum. *American Economic Review* [online], 67 (1), 76–90.
- [74] Skinner, H. and Stephens, P., 2003. Speaking the same language: the relevance of neuro-linguistic programming to effective marketing communications. *Journal of Marketing Communications* [online], 9 (1), 177-192.
- [75] Straub, E., 2009. Understanding Technology Adoption: Theory and Future Directions for Informal Learning. *Review of Educational Research* [online], 79 (2), 625-649.
- [76] The Interaction Design Foundation, 2018. *Gamification* [online]. Denmark: The Interaction Design Foundation. Available from: <https://www.interactiondesign.org/literature/topics/gamification> [Accessed 1 May 2018].
- [77] Thomson Reuters, 2009. China bans physical punishment for Internet addicts [online]. Canada: Thomson Reuters. Available from: <https://in.reuters.com/article/idINIndia-43701020091105> [Accessed 28 June 2018].
- [78] Torous, J. Onnela, J-P. and Keshavan, M., 2017. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Translational Psychiatry* [online], 7 (1), 1-3.
- [79] Turel, O. Serenko, A. and Bontis, N., 2011. Family and work-related consequences of addiction to organizational pervasive technologies. *Information & Management* [online]. 48 (1), 88-95.

- [80] Vaghefi, I. and Lapointe, L., 2014. When Too Much Usage Is Too Much: Exploring the Process of IT Addiction. *Hawaii International Conference on System Science* [online], 47 (1), 4494-4503.
- [81] Venkatesh, V. and Davis, F., 2000. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* [online], 46 (2), 186-204.
- [82] Voss, J. and Paller, K., 2009. An Electrophysiological Signature of Unconscious Recognition Memory. *Nature Neuroscience* [online], 12 (3), 349-355.
- [83] Waelti, P. Dickinson, A. and Schultz, W., 2001. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* [online], 412 (1), 43-48.
- [84] Weinstein, A. and Lejoyeux, M., 2010. Internet Addiction or Excessive Internet Use. *The American Journal of Drug and Alcohol Abuse* [online], 36 (1), 277-283.
- [85] Wessely, S. Bryant, R. Greenberg, N. Earnshaw, M. Sharpley, J. and Hacker Hughes, J., 2008. Does Psychoeducation Help Prevent Post Traumatic Psychological Distress? *Psychiatry* [online], 71 (4), 287-301.
- [86] Yau, Y. and Potenza, M., 2015. Gambling Disorder and Other Behavioral Addictions: Recognition and Treatment. *Harvard Review of Psychiatry* [online], 23 (2), 134-146. United States: Harvard Medical School.
- [87] Yin, H. Ostlund, S. Knowlton, B. Balleine, B., 2005. The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience* [online], 22 (1), 513-523.
- [88] Young, A., 2014. *Brand Media Strategy: Integrated Communications Planning in the Digital Era*. United States: Springer.